Average Treatment Effects in Regression Models With Interactions Between Treatment and Manifest or Latent Covariates

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena

von **Dipl. psych. Felix Flory**
geboren am 20. September 1975 in Heidelberg

Gutachter

1. _____

2. _____

3. _____

Tag des Kolloquiums: _____

Zusammenfassung

In der vorliegenden Dissertation werden statistische Methoden der Datenanalyse zur Untersuchung von bedingten und von durchschnittlichen Behandlungseffekten entwickelt. Dabei werden Interaktionseffekte von Behandlung mit Kovariaten explizit berücksichtig. Die Kovariaten können dabei sowohl manifest als auch latent sein. In vielen Anwendungsfällen werden (durchschnittliche) Behandlungseffekte durch einen einfachen Mittelwertsvergleich der betreffenden Behandlungsgruppen evaluiert. Kovariate werden in Interventionsstudien miteinbezogen, z. B. um die Effekte der Kovariaten selbst oder Interaktionseffekte zu untersuchen, um die Teststärke zu erhöhen oder (in Feldstudien) um eine mögliche kausale Verfälschung zu adjustieren. Die hier dargestellten Verfahren können zum Adjustieren von Mittelwertsunterschieden verwendet werden, wobei es im Unterschied zu bereits existierenden Verfahren möglich ist, Interaktionseffekte von Behandlung und Kovariaten zu berücksichtigen.

Die vorliegende Arbeit beginnt mit einer Definition von durchschnittlichen Effekten. Synonym dazu wird auch der Begriff adjustierter Mittelwertsunterschied verwendet um deutlich zu machen, dass eine kausale Interpretation nicht unbedingt möglich ist. Für manifeste Kovariaten werden

einige der bereits existierenden Verfahren wie ANOVA und ANCOVA genannt und deren Probleme angesprochen. Lösungen in Bezug einiger dieser Problem werden aufgezeigt. Es werden neue Verfahren entwickelt, die eine simultane Analyse von Interaktionseffekten und durchschnittlichen Effekten gestatten. Die Verfahren werden auch dahin erweitert, dass latente Kovariten in die Analyse mitaufgenommen werden können. Monte Carlo Studien werden durchgeführt, die zeigen, dass die entwickelten Verfahren gut für die ausgewählen Beispiele funktionieren. Die entwickelten Verfahren beruhen alle auf dem Prinzip der Maximum-Likelihood-Schätzung, wobei nicht-lineare Bedinungnen auf die Modelparameter bestimmt werden. Es wird ausserdem gezeigt, wie die Teststärke der entwickelten Verfahren geschätzt werden kann.

Abstract

This dissertation develops statistical methods to estimate and test average (or main) treatment effects if treatment effects depend on covariates. The covariates can be manifest or latent. The average effect of a treatment is usually evaluated by comparing the means of the outcome variable between treatment groups. Covariates are included in interventional studies for example, to study covariate effects or interaction effects, to increase power, or (in non-randomized designs) to control for causal bias.

This thesis begins with a definition of *average treatment effects*. The term *adjusted mean difference* is used interchangeably in order to emphasis that the average treatment effect can not be interpreted as the average causal effect of the treatment without making further assumptions. The existing methods for manifest covariates and their problems with regard to testing average treatment effects are discussed. Solutions to some of the problems are provided. New procedures are developed that allow to simultaneously analyze interactions as well as average effects. The procedures are then generalized so that latent covariates can be included in the analysis. Monte Carlo studies show that they work well for the chosen examples. The developed procedures are based on maximum likelihood estimation methods involving non-linear constraints on the model parameters. It is also shown how to estimate power of the outlined procedures with regard to detecting average treatment effects.

Acknowledgments

I am indebted to my supervisor Prof. Dr. Rolf Steyer, who devoted so much time to our discussions about causality, from which I have learned valuable knowledge that will accompany me throughout my career. His suggestions have greatly contributed to the development of the statistical methods described in this dissertation. I wish to thank Prof. Dr. Andreas Klein, University of Western Ontario, for his support. I thank Prof. Dr. Michael Eid, Freie Universität Berlin, for taking the time to review this thesis. All of my reviewers continually stimulated my analytical thinking and greatly assisted me with scientific writing. I am extremely grateful to my parents for their love, their financial support and their encouragement during my studies. My wife Laurie carried me through the time of writing this dissertation and has given me the greatest gift of all. Our daughter Vivienne was born midway of my writings.

Felix J. Flory

Pittsburgh, Pennsylvania, September, 2007

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

In the social sciences, researchers are often interested in the effects of a treatment on an outcome variable (also called dependent variable). A frequently used research design to estimate the effect of the treatment is the treatment-control group design. The observational units are assigned to either the group receiving the treatment (treatment group) or the group receiving no treatment (control group). The effect of the treatment is then estimated by comparing the group means of the outcome.

Covariates are often included in research designs for several reasons; for example, to study the effects of the covariates on the outcome, or to study interaction effects, that is, how the effect of the treatment varies with (or depends on) the covariates. Covariates may simply be included to gain statistical power with regard to detecting treatment effects. Including covariates in research designs where the assignment probabilities of the observational units are not under control by the researcher (so called *observational studies*, see e. g. Rosenbaum, 2002), serves another important purpose, that is, aiming to statistically adjust for bias with regard to the estimation of the treatment effect(s). The goal of these statistical adjustment procedures is

to identify the treatment as the cause of differences between treatment and control group with regard to the estimated outcome.

If interactions between the treatment and the covariates are present, the researcher will surely want to estimate and test hypotheses about these interaction effects. Even in presence of interaction effects the *average effect* or *main effect* of the treatment might provide useful information. This effect will be defined in this dissertation. It can be regarded as the overall effect of the treatment, regardless of the covariates. Estimating and testing this average treatment effect in presence of interaction effects between treatment and covariate is the main focus of this dissertation. This dissertation therefore develops statistical adjustment procedures for the average treatment effects based on models that include interaction effects between treatment and covariates.

For orthogonal factorial designs, ANOVA methods exist that allow to estimate and test hypotheses about average treatment effects as well as interaction effects. There is no dispute about the methods only about whether it is meaningful to interpret average effects in the presence of interactions. This will be a point discussed in this thesis. Another point relates to non-orthogonal designs. The existing ANOVA methods of partitioning the sums of squares are still controversial. Particularly misleading is that the different methods often yield different and sometimes contradictory results (see, e. g. Overall & Spiegel, 1969; Overall, Spiegel, & Cohen, 1975) and the related debate in the Psychological Bulletin.

Also, for studies involving continuous covariates ANCOVA and other multiple linear regression methods have been proposed that are based on centering the covariates in order to analyze average effects. In this dissertation it will be shown that centering the covariate on its mean in advance to es-

timating the model plays an important role with regard to estimating and testing average treatment effects. The reason for this is that the average treatment effect is defined based on the mean of the covariate. It will also be shown however that centering on the covariate mean that is estimated from a sample leads to a biased estimation of the standard error of the average treatment effect.

Methods based on the GLM (including ANOVA and ANCOVA) are often referred to as *multiple linear regression and correlation methods*. However, the maximum likelihood (ML) method described in this dissertation can also be regarded as a multiple linear regression method. Because both methods are compared to each other throughout this dissertation it is necessary to distingish between them. Methods based on the GLM assume fixed regressors whereas the described maximum likelihood methods are more general because they consider stochastic regressors. The values of a fixed regressor are assumed to be fixed, wheres the values of a stochastic regressor are considered to vary at random. The distinction between fixed and stochastic regressors is not important for the estimation and testing of regression coefficients and linear combination of these because GLM methods provide unbiased estimates and tests even for stochastic regressors. The distinction between fixed and stochastic regressors is important however, for the estimation and the test of average effects. This will be shown in the following chapter.

If the effect of the treatment varies with covariates, these variables are also called moderators (see, e. g., Saunders, 1956; Baron & Kenny, 1986). GLM methods (including ANOVA and ANCOVA) are typically used to analyze these interaction effects, involving the estimation and the testing of hypotheses about corresponding regression coefficients or linear combina-

tions of them (see, e. g., Moosbrugger, 1981; Gosslee & Lucas, 1965). Many textbooks on multiple linear regression include these topics (see,e. g., Cohen, Cohen, West, & Aiken, 2003; Keppel & Zedeck, 2000).

If interaction effects are present, a major focus of the analysis will be to estimate and test the conditional effects in order to find out how the effect of the treatment varies depending on the values of the covariates. Nevertheless, a researcher might also be interested in the effect of the treatment overall, that is the main or the average effect of the treatment. For example, in the context of aptitude-treatment interactions, it is frequently the case that individuals at different pre-intervention (baseline) levels on the outcome variable benefit differently from the intervention (see, e. g., Cronbach & Snow, 1977). The question whether the average effect of a treatment differs from zero might be of interest in order to compare different treatments with each other.

Interventional studies often incorporate a treatment-control group design. It is common practice to include covariates that are assumed to influence the outcome variable in order to increase power and to study possible interaction effects. For non-randomized studies it is even more important to include covariates to explain differences between treatment groups at the pre-intervention time point (c. f. Cook & Campbell, 1979).

Among the ANOVA approaches and techniques that have been developed to analyze unbalanced (factorial) designs, four different ways of partitioning the sums of squares are most commonly used and implemented in many statistical software packages. They are often referred to as Type I – Type IV (see, e. g., Searle, 1987; Little, Freund, & Spector, 1991). A problem of these methods is the difficulty to specify the null hypothesis that is tested, especially when interaction effects are present. From a theoretical point of view, none of the hypothesis can be identified as incorrect. They are just

different. But it is usually no easy task to identify the most suitable one for the purposes of a particular study. Different methods may yield different and sometimes contradictory results (see, e. g. Overall & Spiegel, 1969; Overall et al., 1975) and the related debate occurring in the Psychological Bulletin. For a summary see Searle, Casella, and McCulloch (1992) for example.

In the statistics literature the controversy over which are the correct sums of squares in a two-way, unbalanced, ANOVA model including interaction effects is not yet settled. Different perspectives and approaches may be found for example in Snee (1973); M. H. Kutner (1974); Speed, Hocking, and Hacknew (1978). In this thesis, I outline a method that provides a solution to this problem; a definition of average treatment effects is given. Hence, the outlined procedure tests hypotheses that are straightforward to interpret.

ANCOVA methods (see, e. g., Cochran, 1957) have been proposed to include continuous covariates in the analysis of average treatment effects. A major drawback of classical ANCOVA methods is however that possible interaction effects are not considered. Aiken and West (1991) proposed a method that improved the situation considerably. The key is to center the covariates (see also Marquardt, 1980) and use multiple linear regression methods that allow to model interactions. Traditionally, such designs were analyzed using non-optimal adaptations of ANOVA. West, Aiken, and Krull (1996) show how multiple linear regression can produce all of the information provided by traditional but less optimal ANOVA procedures, including an analysis of interaction as well as main (or average) effects. The technique of centering the covariate in advance to the model estimation has also been adopted for the analysis of treatment effects that depend on the *initial status* (i. e. individual differences before the treatment) using latent variable modeling (B. O. Muthén & Curran, 1997).

Based on the theory of causality (Rubin, 1974), the propensity score analysis (Rosenbaum & Rubin, 1984) was developed to analyze average treatment effects specifically for non-randomized studies, also called quasi-experimental designs (see, e. g., Shadish, Cook, & Campbell, 2002) or observational studies (Rosenbaum, 2002). The main principle of this approach is to include covariates that (may) interact with the treatment and on which the treatment assignment might depend on, in order to adjust for bias of the estimated average causal treatment effect. It is obvious that the main focus of this approach is on the average (causal) treatment effect and not on the interaction effects.

The approach proposed in this thesis yields estimates and tests for both: the average treatment effect and the interaction effects. Because the existing multiple linear regression methods are sufficient to estimate and test interaction effects, the focus is on the analysis of average treatment effects. For the case that interactions are present, it is shown that multiple linear regression is only applicable to analyze average treatment effect in rare cases. That is, if the covariate means, that appear in the hypothesis, are known. Because these covariate means have to be estimated in most studies, the maximum likelihood method outlined in this thesis is useful to analyze average treatment effects.

Experimental designs that investigate the effect of a treatment may include covariates in order to study interaction effects and to reduce error variance. Quasi-experimental studies or observational studies may include covariates in order to reduce bias with respect to the estimation of (causal) treatment effects. The maximum likelihood test procedure which will be described and developed for this context is applicable to each type of these studies. Studies based on designs without randomization do not yield estimates of the treatment effects that can be interpreted as the *causal* treat-

ment effect, at least not without making any further assumptions. Therefore, I speak of the average treatment effect (or adjusted mean difference) instead of the average *causal* treatment effect. Whenever appropriate, references are given that point to the theory of causality and necessary assumptions that have to hold in order to interpret the average treatment effect as the average causal treatment effect in the sense of Neyman (1923) or Rubin (1974).

Throughout this thesis I focus on (a single) continuous outcome variable. It is clear, however, that methodology for multiple outcomes, including categorical outcomes, is much needed in practice. Furthermore, the thesis focuses on categorical treatment variables usually considered in intervention studies. It should be pointed out that the proposed methods are applicable to analyze effect of categorical variables in general. Such settings may for example involve gender differences and differences among populations characterized by other categorical variables.

Given that interaction effects are present, it is shown that hypotheses about average treatment effects usually involve the means (i. e. the first moments) of the covariates. If these means are (considered to be) known, then existing methods based on the general linear model (GLM) can be used to test hypotheses about average effects. However, in most cases the covariate means have to be estimated. It will be shown that GLM methods are not the method of choice to analyze average effects if the covariate means have to be estimated. Instead, more general statistical procedures such as the proposed maximum likelihood methods should be used. The simulation study described in the following chapter shows this for a simple design.

Chapter 2 starts with an introductory example illustrating some of the principles of how to define average treatment effects in multiple linear regression analysis. The concept of average treatment effects is then generalized for designs, that include more than one covariate. It is shown how to estimate

and test average treatment effects using the maximum likelihood principle. An example is given that illustrates how to apply the approach to a non-orthogonal design. The role of centering the covariates is also discussed by an example. For each example the implementations in Mplus are provided. A simulation is described showing that maximum likelihood methods are applicable to test average (treatment) effects, if the necessary covariate means have to be estimated. In this case multiple linear regression methods or, to be more precise, methods based on the general linear model, yield inflated type I errors depending on the size of the interaction effects.

The concept of average treatment effects is generalized to the case of latent covariates in chapter 3. Estimating and testing the average treatment effect is then discussed in the succeeding chapters. Chapter 3.3 treats the case in which randomization is successfully implemented in the research design. The proposed solution is based on the general latent variable modeling framework using the multiple group approach. The key is to use constraints that are implied by the randomized design. These constraints facilitate the estimation of the average treatment effect. It is then shown how to compute the power of the proposed approach in order to detect the average treatment effect.

Chapter 4 treats the case in which randomization is not (successfully) implemented in the research design. Research based on such designs is sometimes referred to as non-experimental research, quasi-experiments or observational studies. Two approaches are described for this case and Monte Carlo studies are conducted in order to obtain a first evaluation of both approaches. The Monte Carlo studies show for the given examples that both approaches work well. The conclusion is that the proposed methods succeed in order to estimate and test average treatment effects even in a latent variable framework where the treatment effect varies with latent covariates. This was thought impossible for example by Jaccard and Wan (1996, p. 41).

Chapter 2

Average Effects with Manifest Covariates

2.1 Average Effects

To explain the concept of average effects consider the following simple example. The effects of a treatment on a continuous outcome (or dependent) variable Y are investigated by a study based on a treatment-control group design. The study includes a continuous covariate Z (e.g. a pretest or a personality variable). The group that receives a treatment (indicated with $X = 1$) is compared to a control group ($X = 0$) that receives a different treatment or no treatment at all. The regression equation shall include a (linear) interaction effect between the treatment and the covariate[1]

$$E(Y \mid X, Z) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX. \tag{2.1}$$

[1]The regression models are presented as population models throughout this dissertation. In the social sciences literature regression models are usually represented as sampling models. Equation 2.1 represented as a sampling model (with the usual assumption of i.i.d sampling) for $i = 1, \ldots, N$ observations, can be written as

$$\hat{Y}_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 Z_i X_i.$$

Most readers will be familiar with the multiple regression based approach, often termed *moderated multiple regression* (Aiken & West, 1991; Jaccard, Turrisi, & Wan, 1990; Judd & McClelland, 1989; Saunders, 1956), that is typically used to analyze the data from such a design. It is well known that β_2, the *first-order effect* of X, should in general not be interpreted as the main effect or the average effect of X in the presence of a interaction. A first step to establish the meaning of β_2 is to compute the expected outcome for a given value z of the covariate for each treatment condition, that is

$$\begin{aligned} E(Y \mid X{=}0, Z{=}z) &= \beta_0 + \beta_1 z \\ E(Y \mid X{=}1, Z{=}z) &= \beta_0 + \beta_1 z + \beta_2 + \beta_3 z, \end{aligned} \tag{2.2}$$

for control and treatment group respectively.

The difference of these two values is called the *conditional effect* of the treatment at the value z of the covariate:

$$E(Y \mid X{=}1, Z{=}z) - E(Y \mid X{=}0, Z{=}z) = \beta_2 + \beta_3 z. \tag{2.3}$$

It is apparent that β_2 represents the conditional effect of the treatment at the value 0 of the covariate. Thus, the meaning of the covariate value 0 must be considered in order to interpret β_2. Because most psychological scales rarely have meaningful 0 points, the meaning of β_2 is usually limited.

Hence, centering the covariate ensures that the interpretation of the first-order effect of X will occur at a meaningful value of the covariate. Centering the covariate can be based on $E(Z)$, the true population mean of the covariate, if it is available from census data, for example. Usually however, centering will be based on $\bar{Z}$, the sample mean of the covariate. Because this distinction will be important I denote the centered covariate as $Z' = Z - E(Z)$ or $Z^* = Z - \bar{Z}$. Note that $E(Z)$ is a fixed value but $\bar{Z}$ is a random variable. The values $\bar{z}$ of this random variable change randomly from sample to sample.

Based on Z', the regression of Equation 2.1 can be written as

$$E(Y \mid X, Z) = \beta'_0 + \beta'_1 Z' + \beta'_2 X + \beta'_3 Z' X \tag{2.4}$$

with

$$\beta'_0 = \beta_0 + \beta_1 E(Z) \qquad \beta'_1 = \beta_1 \tag{2.5}$$

$$\beta'_2 = \beta_2 + \beta_3 E(Z) \qquad \beta'_3 = \beta_3. \tag{2.6}$$

For a given value of $\bar{Z} = \bar{z}$, the covariate centered on the sample mean is $Z^* = Z - \bar{z}$, and the corresponding regression may be written as

$$E(Y \mid X, Z) = \beta^*_0 + \beta^*_1 Z^* + \beta^*_2 X + \beta^*_3 Z^* X \tag{2.7}$$

with

$$\beta^*_0 = \beta_0 + \beta_1 \bar{z} \qquad \beta^*_1 = \beta_1 \tag{2.8}$$

$$\beta^*_2 = \beta_2 + \beta_3 \bar{z} \qquad \beta^*_3 = \beta_3. \tag{2.9}$$

If the regression is based on the centered covariate, then β'_2 represents the conditional effect of the treatment at the true population mean; whereas, β^*_2 represents the conditional effect of X at the sample mean of the covariate. In both cases, centering the covariate ensures that the interpretation of the first-order effect of X will occur at a meaningful value of the covariate (see, e. g., West et al., 1996).

It can also be seen that the first-order effect of the covariate and the interaction effect remain the same whether centering is applied or not. The fact that the highest-order effect remains constant across any linear rescaling of the covariate is well known, whereas other effects generally vary. Methods based on the GLM can be used to test hypotheses about the regression coefficients. I note in passing that the corresponding t-test of β_3 yields constant results across any linear rescaling of Z.

The question is, if the first-order effect of the treatment in the regressions based on the centered covariate (i. e. β_2' and β_2^*) can be interpreted as the average treatment effect. To answer this question, I first define the term average effect of the treatment. Consider the function $g_{1-0}(Z)$ that maps each covariate value to the conditional treatment effect

$$g_{1-0}(Z) = \beta_2 + \beta_3 Z. \tag{2.10}$$

This function is called the *effect function* (see, e. g. Steyer, Partchev, Kröhne, Nagengast, & Fliege, 2007). The covariate is called a *modifier*. Synonyms for modifier and effect function are moderator and moderator function (see, e. g., Saunders, 1956; Baron & Kenny, 1986). The effect function of the given example shows that the effect of the treatment depends (linearly) on the covariate.

The average effect of the treatment (with regard to the covariate) for the regression of the given example is defined simply as the average of the conditional effects. To be more precise, the *average effect* (AE) of treatment 1 vs. treatment 0 on the (expected value of the) outcome variable Y for the given regression is the expectation of the effect function $g_{1-0}(Z)$:

$$AE_{1-0} = E\left(g_{1-0}(Z)\right) = E\left(\beta_2 + \beta_3 Z\right) = \beta_2 + \beta_3\, E(Z). \tag{2.11}$$

Note that the average effect is defined with regard to the given regression including the covariate Z. For a regression including a different covariate the average effect might differ. However, if the regression of Equation 2.1 is causally unbiased (e. g. given a randomized design, or conditional randomization on Z), then it can be shown that the average treatment effect is the same for all (combinations of) possible covariates. Given a non-randomized design, the regression is causally unbiased for example if Z is the only confounding variable. The interested reader is referred to literature on causality

(e. g. Rubin, 2006). The concept of the average effect as it is defined here is also described by Angrist, Imbens, and Rubin (1996) in a similar framework. Rosenbaum (2002, pp. 77–78) describes a method to estimated this average treatment effect and calls this method *adjustment for overt bias*. Steyer, Gabler, von Davier, and Nachtigall (2000) give an in depth discussion on sufficient conditions implying a regression to be causally unbiased (see also Steyer et al., 2007; Steyer, Nachtigall, Wüthrich-Martone, & Kraus, 2002).

The answer to the question from above can now be given. Distinguished are two cases: first, the centering is based on the $E(Z)$, the population mean of the covariate, and second, the centering is based on $\bar{Z}$, the sample mean of the covariate.

Centering on the population mean: Comparing the definition of the average effect of Equation 2.11 with the meaning of β_2' in Equation 2.6 reveals that the first-order effect of X is equivalent to the average effect. Consequently, the average treatment effect can be estimated by centering the covariate on the population mean and estimating β_2' using multiple linear regression. The corresponding t-test of $\beta_2' = 0$ can be used to test the *null hypothesis*

$$H_0 : \ AE_{1-0} = \beta_2 + \beta_3 \, E(Z) = 0, \tag{2.12}$$

stating that no average treatment effect is present.

This test can also be performed with an R^2-difference test comparing the full model against the model without the first-order effect of X. The main limitation of this approach is that the population mean of the covariate has to be known in order to perform the centering. As mentioned before, the population mean of the covariate might be inferred from some census data.

Centering on the sample mean: Comparing the definition of the average effect of Equation 2.11 with the meaning of β_2^* in Equation 2.9 reveals

that the first-order effect of X is only equivalent if $\bar{z} = E(Z)$ (given a non-zero interaction effect). However, a sample mean $\bar{z}$ that is used to compute the centered covariate will almost always differ from the true population mean $E(Z)$. Hence, β_2^* will almost always differ from the average effect. The simulation study described later clearly shows the consequences. A multiple linear regression estimate for β_2^* still yields an unbiased estimate of the average treatment effect. Yet, the t-test of $\beta_2^* = 0$ (as well as the equivalent R^2-test) yields inflated type I errors with regard to testing the average treatment effect (depending on β_3).

The reason for this can be explained with the hypothesis that is tested. The t-test of $\beta_2^* = 0$ does exactly what it is supposed to do. Considering Equation 2.9 it is testing the hypothesis $\beta_2^* = \beta_2 + \beta_3\bar{z} = 0$. However, this hypothesis differs from the null hypothesis given in Equation 2.12 depending on $\bar{Z} = \bar{z}$ (and β_3). Because the simulation study was designed so that H_0 holds, the hypothesis $\beta_2^* = \beta_2 + \beta_3\bar{z} = 0$ almost never holds (given $\beta_3 \neq 0$). The hypothesis actually changes depending on the sample and its mean $\bar{z}$. Testing a hypothesis that does not hold leads to a probability of rejecting the hypothesis that is larger than the significance level. This can be seen in the simulation study. Hence, it is inappropriate to test H_0 with a t-test of $\beta_2^* = 0$, because this test yields inflated type I errors depending on β_3.

So far we have seen that multiple linear regression should only be used if the centering of the covariate is based on the population mean. Because the mean of the covariate will rarely be known, a method to test the average treatment effect is needed if the covariate mean can only be estimated from the sample. Before I propose a solution to this problem I describe a multiple linear regression test based on the linear hypothesis that is equivalent to the centering approach for the given example, yet more flexible as we will see in

an example later.

Hypotheses about (linear combinations of) regression coefficients can be tested with the *linear hypothesis*

$$\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{\delta} = \mathbf{0} \tag{2.13}$$

by specifying a hypothesis matrix $\boldsymbol{A}$ and a vector $\boldsymbol{\delta}$ according to $\boldsymbol{\beta}$, the vector of regression coefficients. The corresponding F-test is described in many textbooks on multiple regression (see, e. g., Fox, 1997; Searle, 1971). Computer programs for multiple linear regression[2] (e. g., R, SAS, S-Plus) that allow testing linear hypotheses can be used to perform this test.

For the given example, hypotheses about average treatment effects can be tested with the hypothesis matrix

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 1 & E(Z) \end{pmatrix} \tag{2.14}$$

against any value of $\boldsymbol{\delta}$. Setting $\boldsymbol{\delta} = 0$ specifies the hypothesis of no average treatment effect and the corresponding F-test yields exactly the same results as centering the covariate on the population mean and testing β_2' as described above. The advantage of the centering approach is of course that no hypothesis matrix has to be specified. The linear hypothesis however, provides more flexibility. We will later see for example, that the linear hypothesis test can be applied to designs that include interactions between continuous covariates.

The main limitation is again that $E(Z)$ has to be known in order to specify the hypothesis matrix. Usually however, only the sample mean $\bar{Z}$ will be available. For a given sample with $\bar{Z} = \bar{z}$ the linear hypothesis test uses

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 1 & \bar{z} \end{pmatrix}. \tag{2.15}$$

[2]It is possible to perform this test with SPSS using the scripting language.

Because $\bar{z}$ will deviate to some extent from $E(Z)$, the hypothesis that is tested is not equivalent to the null hypothesis H_0.

The multiple regression methods (centering and the linear hypothesis) based on the sample mean may lead to potential problems with regard to the estimation and the testing of hypothesis about average treatment effects. $\bar{z}$ is treated as a fixed value and not as an estimate for the (unknown) covariate mean $E(Z)$. In other words, the null hypothesis H_0 combines two (unknown) terms β_3 and $E(Z)$ together in a multiplicative way yielding a non-linear hypothesis. Hence, a statistical test is needed that eliminates the given limitation of the GLM.

The following outline proposes a maximum likelihood test that involves non-linear constraints in order to test the null hypothesis H_0. For the given example the unconstrained model of Equation 2.1 is compared to a model with the non-linear constraint

$$\beta_2 = -\beta_3 \, E(Z). \tag{2.16}$$

The chi-square values of both models are compared by a chi-square difference test. The implementation in Mplus of the test for the given example is provided after the general outline of the maximum-likelihood procedure. Before I describe the general outline of the maximum likelihood test, I would like to address a question that some readers might raise at this point.

2.2 Simple Group Mean Differences

With regard to the specification of the average treatment effect, some readers might suggest to take the difference between the mean of Y in the treatment group and compare it to the mean of Y in the control group (and ignore the covariate). Given a randomized design this would lead to a causally unbiased

estimation of the average treatment effect. However, the power of this test would be less if outcome variance is explained by the covariate. Given a non-randomized design the mean difference $E(Y \mid X{=}1) - E(Y \mid X{=}0)$ might not always be equal to the average treatment effect, a numerical example that is based on the example above is given in the following section.

In a non-randomized design it might well be that the two treatment populations differ with respect to their covariate means. Let the covariate mean of the control population be $E(Z \mid X{=}0) = 5.4$ and the covariate mean of the treatment population be $E(Z \mid X{=}1) = 5.0$. For simplicity, let the samples of each population be of equal size so that the expected value of the covariate results in $E(Z) = 5.2$. The regression coefficients of Equation 2.1 shall be $\beta_0 = -8.6$, $\beta_1 = 1.8$, $\beta_2 = -1.78$ and $\beta_3 = 0.4$.

Based on Equation 2.11 the computation of the average treatment effect yields a value of 0.3. The means of Y for the two groups can be computed as[3]

$$\begin{aligned} E(Y \mid X{=}0) &= \beta_0 + \beta_1 \, E(Z \mid X{=}0) \\ &= -8.96 + 1.8 \cdot 5.4 = 0.76 \end{aligned} \tag{2.17}$$

$$\begin{aligned} E(Y \mid X{=}1) &= \beta_0 + \beta_1 \, E(Z \mid X{=}1) + \beta_2 + \beta_3 \, E(Z \mid X{=}1) \\ &= -8.96 + 1.8 \cdot 5 - 1.78 + 0.4 \cdot 5 = 0.26 \end{aligned} \tag{2.18}$$

The resulting mean difference of $E(Y \mid X{=}1) - E(Y \mid X{=}0) = -0.5$ differs considerably from average treatment effect of 0.3. The sign actually changes. Holland (1986) called this mean difference the *prima facie effect*. The fact that the prima facie effect can be misleading is well known (see, e. g., Simpson, 1951). In the given example the prima facie effect differs from the average treatment effect because the groups differ with regard to the covariate, which itself affects the outcome.

[3]See the Appendix C.1 for a detailed description about the calculation.

In a randomized experiment the (population) means of the covariate of the two groups will be equal, i. e. $E(Z \mid X{=}0) = E(Z \mid X{=}1) = E(Z)$. As a consequence, the mean difference

$$E(Y \mid X{=}1) - E(Y \mid X{=}0) = \beta_2 + \beta_3 E(Z) \tag{2.19}$$

will be the same as the average treatment effect. Given a successful randomization the average effect can be tested by testing this simple mean difference. It should be pointed out though that the loss of power might be considerable if the ignored covariate has a strong effect on the treatment. Including such a covariate (if available) in the analysis of average treatment effect might therefore be indicated even if randomization was implemented in the design of a study.

For the analysis of treatment effects from non-randomized studies it is even more important to include covariates that explain differences in the treatment groups. If the treatment groups differ with respect to some covariates as in the last example, and these covariates affect the dependent variable, then these covariates should be included in the analysis. This adjustment technique is well known in statistical literature and sometimes referred to as *(statistically) controlling for covariates*, or *partialling out the effects of the covariates*. This example was given in order to emphasize the distinction between the average treatment effect and the simple group mean difference.

2.3 Generalizations

The example of the last section included only one (continuous) covariate and only two groups. In the following section, the concept of the average treatment effect is generalized for more than two treatments with possibly multiple (continuous and categorical) covariates.

Let there be J treatment groups. The treatment variable X takes on the value $X = 1$ for control, and $X = 2, \ldots, X = J$ for the treatment groups. The following $J - 1$ dummy variables are used to identify each treatment: $I_{X=2}, \ldots, I_{X=J}$ indicate with 1 and 0 whether or not the observational unit is assigned to treatment j, with $j = 2, \ldots, J$. The covariates are summarized in the vector $\boldsymbol{Z}$, which needs to be specified in each application.

The regression equation of the outcome variable Y regressed on the treatment and the covariates can generally be written as

$$E(Y \mid X, \boldsymbol{Z}) = g_1(\boldsymbol{Z}) + g_{2-1}(\boldsymbol{Z}) \cdot I_{X=2} + \cdots + g_{J-1}(\boldsymbol{Z}) \cdot I_{X=J}. \qquad (2.20)$$

The function g_1 is called the *ordinate function*. It can be regarded as a random variable that maps each (combination of) covariate value(s) to $E_{X=1}(Y \mid \boldsymbol{Z})$, the expected outcome under control. The interpretation of the functions $g_{2-1}, \ldots, g_{J-1}$ is straightforward. Each function $g_{j-1}(\boldsymbol{Z}) = E_{X=j}(Y \mid \boldsymbol{Z}) - E_{X=1}(Y \mid \boldsymbol{Z})$ can be regarded as a random variable that maps every (combination of) covariate value(s) to the conditional effect of the treatment j given $Z = z$, which is the difference between the expected outcome under treatment j and the expected outcome under control. These functions are called *slope* or *effect functions*.

A parametrization of the regression (the ordinate and the effect functions) is needed for a statistical analysis. Consider the following linear parametrization

$$E(Y \mid X, \boldsymbol{Z}) = \boldsymbol{z}'\boldsymbol{\gamma}_1 + \boldsymbol{z}'\boldsymbol{\gamma}_2 I_{X=2} + \cdots + \boldsymbol{z}'\boldsymbol{\gamma}_J I_{X=J}, \qquad (2.21)$$

which is used in multiple linear regression.

The vector $\boldsymbol{z}' = (\,1 \;\; Z_1 \;\; Z_2 \;\; \ldots\,)$ consists of random variables constructed from the covariates. The first entry might be dropped if the ordinate and effect functions do not include an intercept. The components of $\boldsymbol{z}$ may represent continuous covariates and code (e. g., dummy) variables of categorical

covariates. Some components might represent products of covariates (or of their code variables), whereas others might be created by (different) functions of the covariates: covariates raised to a power, and products of these. The vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J$ contain the corresponding regression coefficients. This parametrization is well documented in the literature on multiple linear regression (see, e. g., Cohen et al., 2003; Draper & Smith, 1981; M. Kutner, Nachtsheim, & Neter, 2004).

To give an example: the regression of Equation 2.1 can be expressed with $\boldsymbol{z}' = (1 \;\; Z)$, $\boldsymbol{\gamma}_1 = (\beta_0 \;\; \beta_1)'$, and $\boldsymbol{\gamma}_2 = (\beta_3 \;\; \beta_4)'$. The example on centering later in this dissertation includes an interaction between two continuous covariates, i. e.

$$\boldsymbol{z}' = \begin{pmatrix} 1 & Z_1 & Z_2 & Z_1 Z_2 \end{pmatrix}. \tag{2.22}$$

2.3.1 Average Effects

Above, it was mentioned that each effect function $g_{j-1}(\boldsymbol{Z})$ (with $j = 2, \ldots, J$) can be regarded as a random variable that maps each (combination of) covariate value(s) to the conditional effect of the treatment j. The following definition of average effects is therefore straightforward.

Definition 1. *The average effect AE_j of treatment j with regard to the control group $X = 1$ and the covariate $\boldsymbol{Z}$ is defined as the average of the effect function $g_{j-1}(\boldsymbol{Z})$, i. e.*

$$AE_{j-1} = E\left(g_{j-1}(\boldsymbol{Z})\right). \tag{2.23}$$

Note that the concept of the average effects is not limited to a linear parametrization that is used in multiple linear regression. If there is no interaction between a treatment j and the covariates, the function $g_{j-1}(\boldsymbol{Z})$ will be a constant. In other words the effect of the treatment j on the outcome

variable is the same for each (combination of) covariate value(s). It is trivial that in this case the average effect is equal to this constant effect, and the model is equivalent to a traditional ANCOVA model.

The definition of the average effect of the treatment is given with regard to the covariates included in the regression of Equation 2.20. Changing the (combination of the) covariates might very well lead to different average treatment effects unless the regression in Equation 2.20 is causally unbiased. As mentioned before the term *average effect* does not mean that it can be interpreted as the *average causal effect* in the sense of Rubin (1974) without further assumptions. In a randomized experiment, as well as in an experiment with conditional randomization based on the covariates included in the regression, the average effect is equal to Rubin's average causal effect. For conditions that have to hold so that the average effect can be interpreted as the average causal effect outside of (conditional) randomized experiments, the interested reader is referred to the literature on causality (see, e. g., Gelman & Meng, 2004; Rubin, 1974; Rubin, 1978; Rubin, 2006; Steyer et al., 2002).

2.4 Estimating and Testing Average Effects

Multiple linear regression can be used to estimate and test hypotheses about the regression coefficients $\gamma_1, \ldots, \gamma_J$ and compute their confidence intervals. Interaction effects can be estimated and tested for significance, revealing detailed information about how a treatment affects the outcome for every value of the covariates.

For the average effects things are more complicated. Given the linear parametrization of Equation 2.21 it is straightforward to compute the average

effect of treatment j as

$$E\left[g_{j-1}(\boldsymbol{Z})\right] = E\left(\boldsymbol{z}'\boldsymbol{\gamma}_j\right) = \begin{pmatrix} 1 & E(Z_1) & E(Z_2) & \dots \end{pmatrix}' \boldsymbol{\gamma}_j \,. \tag{2.24}$$

Within the multiple linear regression framework (i. e. the general linear model), the linear hypothesis provides a way to estimate and test hypotheses about the average effects. By specifying a linear hypothesis matrix according to Equation 2.24, hypotheses about the average effects can be tested using multiple linear regression. An example for a linear hypothesis, the corresponding hypothesis matrix, and references were given in the introductory example.

The main restriction is that the elements $E(Z_1), E(Z_2), \dots$ of the hypothesis matrix have to be known. They might be considered to be known for example, if the whole population is included in the study, or the means are inferred from some census data or from previous surveys that used quota samples trying to perfectly reproduce the underlying population. The population mean of the covariate is also known if the covariate represents a second experimental factor in the research design. In this case the researcher has control of the covariate values. Another way to distinguish between whether or not multiple linear regression is applicable is to ask the following question: Would the values in the hypothesis matrix be the same or are they likely to change (due to the sampling process) in a replication study? If the values in the hypothesis matrix would be the same in replication studies, then multiple linear regression is applicable to test the average treatment effect.

Usually however, the means $E(Z_1), E(Z_2), \dots$ have to be estimated, and therefore become parameters in the underlying statistical model. Consequently, hypotheses about average effects are not linear. The hypotheses are non-linear because they include products of model parameters: $E(Z_1) \cdot \gamma_{j1}, E(Z_2) \cdot \gamma_{j2}, \dots$. For a simple setting the simulation study described later

shows that multiple linear regression methods (based on the general linear model) yield inflated probabilities of Type I errors: the larger the interactions between treatment and covariates, the more inflated the probability of a Type I error.

2.4.1 Maximum Likelihood Methods

If the means $E(Z_1), E(Z_2), \ldots$ of the covariates, that appear in hypotheses about average effects (see Equation 2.24) have to be estimated, maximum likelihood methods can be used to test these (non-linear) hypotheses. This is usually the case if the covariates are stochastic regressors. The values of these covariates are not fixed (by the experimenter) but are observed as they randomly occur and population means of these covariates are not available. The next paragraphs give an outline of the test.

Consider the following random experiment: $i = 1, \ldots, N$ observations are drawn from a population. For each observation the values of the following random variables are recorded: $\boldsymbol{z}$ a vector of covariates (as well as numerical functions of the covariates), the treatment variable X and the outcome variable Y. The following assumptions are made:

- $\boldsymbol{z}$ has a parametric distribution with a set ρ_Z of parameters (e. g., if $\boldsymbol{z}$ is distributed multivariate normally, then the parameter set consists of the vector of means and the variance-covariance matrix of $\boldsymbol{z}$).

- The conditional distribution of X given $\boldsymbol{z}$ is parametric with a set ρ_X of parameters.

- The conditional distribution of Y given X and $\boldsymbol{z}$ is parametric with a set ρ_Y of parameters.

The probability of a single observation $p_{\rho_X,\rho_Y,\rho_Z}(x_i, y_i, \boldsymbol{z}_i)$ can be decomposed into: $f_{\rho_Z}(\boldsymbol{z}_i)$, the probability of the covariate vector; $g_{\rho_X}(x_i \mid \boldsymbol{z}_i)$, the conditional probability of X given the covariate vector; and $h_{\rho_Y}(y_i \mid x_i, \boldsymbol{z}_i)$, the conditional probability of Y given treatment and the covariate vector.

If the N observations can be regraded as different independent random trials, the likelihood of the data is the product of the likelihood of all observations. The logarithm of this likelihood (log-likelihood) can then be written as

$$\begin{aligned}\mathrm{logL}(\rho_X, \rho_Y, \rho_Z) = &\sum_{i=1}^{N} \log\big(f_{\rho_Z}(\boldsymbol{z}_i)\big) \\ &+ \sum_{i=1}^{N} \log\big(g_{\rho_X}(x_i \mid \boldsymbol{z}_i)\big) + \sum_{i=1}^{N} \log\big(h_{\rho_Y}(y_i \mid x_i, \boldsymbol{z}_i)\big). \quad (2.25)\end{aligned}$$

Because the summands are functions of mutually exclusive sets of parameters the maximum of the log-likelihood is

$$\begin{aligned}\max_{\rho_X,\rho_Y,\rho_Z} \big[\mathrm{logL}(\rho_X, \rho_Y, \rho_Z)\big] = &\max_{\rho_Z}\left[\sum_{i=1}^{N} \log\big(f_{\rho_Z}(\boldsymbol{z}_i)\big)\right] \\ + \max_{\rho_X}&\left[\sum_{i=1}^{N} \log\big(g_{\rho_X}(x_i \mid \boldsymbol{z}_i)\big)\right] + \max_{\rho_Y}\left[\sum_{i=1}^{N} \log\big(h_{\rho_Y}(y_i \mid x_i, \boldsymbol{z}_i)\big)\right]. \quad (2.26)\end{aligned}$$

The set of values $\hat{\rho}_X, \hat{\rho}_Y, \hat{\rho}_Z$, that maximize the log-likelihood are taken as the estimates for the (unknown) parameters ρ_X, ρ_Y, ρ_Z.

A hypothesis about average effects poses a constraint on the parameter sets ρ_Y, ρ_Z (see Equation 2.24), because it includes parameters of the covariate distribution and parameters of the conditional distribution of Y given X and $\boldsymbol{z}$ (some of the regression coefficients). It is therefore possible that maximizing $\sum \log(f_{\rho_Z}(\boldsymbol{z}_i))$ leads to suboptimal values of $\sum \log(h_{\rho_Y}(y_i \mid x_i, \boldsymbol{z}_i))$ and vice versa. Hence, the set of values $\hat{\rho}'_Z, \hat{\rho}'_Y$ that maximize the likelihood under the constraint will usually differ from the set of values $\hat{\rho}_Z, \hat{\rho}_Y$ that maximize the log-likelihood without a constraint.

However, the choice of the parameter set ρ_X has neither an effect on $\sum \log(f_{\rho_Z}(\boldsymbol{z}_i))$, nor on $\sum \log(h_{\rho_Y}(y_i \,|\, x_i, \boldsymbol{z}_i))$. It also does not depend on the choice of ρ_X, whether the constraint is fulfilled or not. Consequently, the constraint can be ignored in order to maximize $\sum \log(g_{\rho_X}(x_i \,|\, \boldsymbol{z}_i))$. This implies $\hat{\rho}'_X = \hat{\rho}_X$, which means the values that maximize $\sum \log_C(g_{\rho_X}(x_i \,|\, \boldsymbol{z}_i))$ under the constraint are the same as in the unconstrained case. Hence, the maximum of the likelihood under the constraint can be written as

$$\begin{aligned} \max_{\rho_X,\rho_Y,\rho_Z} [\log \mathrm{L}_C(\rho_X, \rho_Y, \rho_Z)] &= \max_{\rho_X} \left[\sum_{i=1}^{N} \log \left(g_{\rho_X}(x_i \,|\, z_i) \right) \right] \\ &+ \max_{\rho_Y,\rho_Z} \left[\sum_{i=1}^{N} \log_C \left(f_{\rho_Z}(z_i) \right) + \sum_{i=1}^{N} \log_C \left(h_{\rho_Y}(y_i \,|\, x_i, z_i) \right) \right]. \end{aligned} \tag{2.27}$$

If a hypothesis about average effects is true, then a chi-square distributed test statistic results by taking twice the difference between the maximum of the likelihood of the constrained model and the maximum of likelihood of the unconstrained model:

$$\begin{aligned} \chi^2 &:= 2 \left\{ \max_{\rho_X,\rho_Y,\rho_Z} \left(\log \mathrm{L}(\rho_X, \rho_Y, \rho_Z) \right) - \max_{\rho_X,\rho_Y,\rho_Z} \left(\log \mathrm{L}_C(\rho_X, \rho_Y, \rho_Z) \right) \right\} \\ &= 2 \left\{ \max_{\rho_Z} \left(\sum_{i=1}^{N} \log \big[f_{\rho_Z}(z_i) \big] \right) + \max_{\rho_Y} \left(\sum_{i=1}^{N} \log \big[h_{\rho_Y}(y_i \,|\, x_i, z_i) \big] \right) \right. \\ &\qquad \left. - \max_{\rho_Y,\rho_Z} \left(\sum_{i=1}^{N} \log_C \big[f_{\rho_Z}(z_i) \big] + \sum_{i=1}^{N} \log_C \big[h_{\rho_Y}(y_i \,|\, x_i, z_i) \big] \right) \right\}. \end{aligned} \tag{2.28}$$

The degrees of freedom equal the number of average effects occurring in the hypothesis.

This outline shows that it is possible to test average effects even if the parameters of the covariate that appear in the hypothesis have to be estimated. The test is based on the maximum likelihood principle. The maximum likelihood of an unconstrained model is compared to the maximum likelihood of a

model that has a non-linear constraint posed on some parameters according to the null hypothesis. The key is that the conditional distribution of X given the covariates does not have to be considered, which makes it possible to conduct the test with programs like LISREL (Jöreskog & Sörbom, 1996) and Mplus (L. Muthén & Muthén, 2004) that allow maximum likelihood methods involving non-linear constraints.

2.4.2 Implementation in Existing Software

Using a program such as LISREL or Mplus will result in additional assumptions about the distribution of the covariates and the conditional distribution of Y given X and the covariates. Two models are specified according to the regression equation — an unconstrained model and a model with (non-linear) constraints according to the null hypothesis. Instead of taking the difference between maximum log-likelihood values of the two models, it is also possible to take the difference between the chi-square values of the two models. Both differences yield equivalent p values because the chi-square values are computed using the same baseline model (i. e. the corresponding saturated model with zero degrees of freedom). The unrestricted model might sometimes even be the baseline model, so that the respective chi-square value will be zero. In this case, the chi-square value of the restricted model can directly be taken as the test statistic, and the p value of the restricted model is already the p value according to the null hypothesis. If the unrestricted model is not equal to the baseline model, the difference in the chi-square values is chi-square distributed with the degrees of freedom equal to the number of constraints, and the corresponding p value has to be computed.

Based on the results from the unconstrained model, it is possible to compute point estimates for the average effects and their standard errors. These

standard errors can either be requested directly from the program, or they can be computed using the multivariate delta method (Cox & Hinkley, 1974; Stuard & Ord, 1994) (the necessary variance-covariance matrix of the involved parameters can be requested from the program).

The procedure above provides not only a test of hypotheses about average effects. The results from the unconstrained model also provide the estimates and tests of the regression coefficients and therefore the information about the interaction effects and the conditional effects. In this way, the outlined procedure provides an in depth view about how treatments work, yielding both an analysis of the interaction effects and an analysis of the average treatment effects. The resulting information is similar to the information that is provided by an orthogonal design analyzed with ANOVA. However, the procedure is applicable to a much wider range of designs: non-orthogonal designs and designs including continuous covariates.

2.4.3 Power Estimation of Average Treatment Effects

A measure of effect size is needed in order to analyze power with regard to average treatment effects. In a traditional two-group t-test setting, effect size is typically defined as the treatment and control group difference in outcome means, divided by a standard deviation based on the pooled outcome variance (Cohen, 1988). As mentioned before, this (simple) outcome mean difference can be misleading in non-randomized studies with interaction effects between treatment and covariates.

It is straightforward to define an effect size for the average treatment effect that is in line with Cohen: the average treatment effect divided by the standard deviation of the outcome variable. A possible variation is to divide the average treatment effect by the standard deviation of the outcome of the

control group. In this way, the control group provides the normative value, whereas the treatment group variance in part reflects the treatment effect.

Given a measure of effect size, power can in principle be estimated by carrying out a Monte Carlo study recording the proportion of replications in which the incorrect model is rejected. However, Satorra and Saris (1985) proposed a method that provides a major simplification. This method is especially suitable to estimate the power to detect average treatment effects, because a hypothesis about the absence of an average treatment effect can be viewed as a specific misspecification of the model (see also Saris & Satorra, 1993; Saris & Stronkhorst, 1984). This method is described in detail in section 3.5.

Because the outline of the procedure was done in a general way, the following paragraphs provide examples that show how to apply the procedure using Mplus in order to test average effects. The examples include the introductory example, an example of a non-orthogonal design, and an example that illustrates centering the covariates.

2.4.4 The Test for the Introductory Example

The introductory example included a continuous covariate as well as a treatment and a control group. The regression of Equation 2.1 is repeated here:

$$E(Y \mid X, Z) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX. \tag{2.29}$$

The null hypothesis H_0 stating that no average treatment effect is present was given in Equation 2.12, which is repeated here:

$$H_0 : \; AE_{1-0} = \beta_2 + \beta_3 \, E(Z) = 0 \tag{2.30}$$

If $E(Z)$ is known, it was shown how multiple linear regression can be used to test the null hypothesis. It will however often be the case that $E(Z)$ has to

```
1  DATA: FILE IS data.dat;
2  VARIABLE: NAMES ARE Y Z X;
3  USEVARIABLES ARE Y Z X ZX;
4  DEFINE: ZX = Z * X;
5  ANALYSIS: TYPE = MEANSTRUCTURE;
6  MODEL:
7  Y ON Z(b1)
8       X(b2)
9      ZX(b3);
10 [Y X ZX];
11 [Z](mZ);
12 MODEL CONSTRAINT: b2 = - b3 * mZ;
```

Listing 2.1: Mplus input for the constrained model of the introductory example

be estimated and the outlined maximum likelihood procedure has to be used in order to test the hypothesis. The restricted model with the constraint given in Equation 2.16 is tested against the unrestricted model using chi-square difference testing. For the given example, the unrestricted model is identical to the saturated (or baseline) model that Mplus uses to compute the chi-square statistic, resulting in a value of zero. Therefore, the chi-square value of the restricted model directly yields the correct test statistic and the p value of the output is the probability of the test statistic under the null hypothesis. The Mplus input for the restricted model is given in Listing 2.1. The unrestricted model is simply obtained by deleting line 12.

An estimate for the average effect can be computed by applying Equation 2.11 using parameter estimates from the unrestricted model. The corresponding standard error estimate can either be requested from Mplus directly or it can be computed using the multivariate delta method (Cox & Hink-

ley, 1974; Stuard & Ord, 1994; Wasserman, 2004). The necessary variance-covariance matrix of the parameters can be requested from Mplus [4].

2.5 ANOVA

A key feature of the described maximum likelihood method is to solve a problem that has puzzled statisticians for decades: the analysis of non-orthogonal (or in general unbalanced) designs (see, e. g., Carlson & Timm, 1974; Gosslee & Lucas, 1965; Keren & Lewis, 1976). ANOVA is the technique that is most commonly used to analyze such designs. Because it is not possible for non-orthogonal designs to partition the sums of squares as a sum of the factors, interactions, and error sums of squares, many attempts were undertaken to find an adequate partitioning (see Searle et al., 1992, for a summary).

Among the approaches and techniques that have been developed, four different ways of partitioning the sums of squares are commonly used and implemented in many statistical software packages. They are usually referred to as Type I – Type IV (see, e. g., Searle, 1987; Little et al., 1991). The major problem of these methods is often to clearly specify the null hypothesis that is tested, especially when interaction effects are present. Sometimes the results may even be contradictory (Overall & Spiegel, 1969; Overall et al., 1975). It can also be shown that these four types of partitioning the sums of squares generally do not test the hypothesis of no average effect as specified in this dissertation (see Wüthrich-Martone, 2001, for an example).

Data from non-orthogonal designs may include a (categorical) treatment variable and some categorical covariates (also called factors) that predict the outcome variable. The treatment variable and the covariates have to be

[4] See the `TECH3` option in the manual (L. Muthén & Muthén, 2004)

coded in an appropriate coding system (see, e. g., Cohen et al., 2003) in order to apply the outlined method. Dummy coding will be used throughout this dissertation.

The linear hypothesis of the multiple linear regression approach (i. e. the general linear model) can be used to test hypotheses about average effects, in case the covariate means appearing in the hypotheses are known. If these means have to be estimated from the sample, the maximum likelihood procedure has to be applied to test these (non-linear) hypotheses.

In this manner the outlined procedure overcomes the problems of the non-orthogonal ANOVA. According to Definition 1 (and Equation 2.24), the hypotheses that are being tested are straightforward and do not change depending on the cell frequencies. The following section gives an example how to test the average treatment effects for a simple factorial design.

2.5.1 A Simple Factorial Design

Consider a hypothetical 3×3 factorial design. The factor X represents three treatment groups. The factor Z represents a categorical covariate with three levels. The cell frequencies may vary arbitrarily. It actually matters little to the outlined procedure whether the design is orthogonal or non-orthogonal, as long as there are sufficient observations in each cell. The treatment variable X may be dummy coded with $I_{X=1} = 1$ for the first group and 0 else; $I_{X=2} = 1$ for the second group and 0 else, consequently the third group serves as the reference. The covariates may also be dummy coded with $Z_1 = 1$ for level one and zero else, and $Z_2 = 1$ for level two and zero else, so that level three serves as the reference[5]. The regression equation

[5]Dummy coding was chosen for simplicity. However, every appropriate coding scheme can be applied. The resulting computation will differ.

can be written in the following way:

$$\begin{aligned} E(Y \mid X, Z) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 I_{X=1} + \beta_4 I_{X=2} \\ + \beta_5 Z_1 I_{X=1} + \beta_6 Z_2 I_{X=1} + \beta_7 Z_1 I_{X=2} + \beta_8 Z_2 I_{X=2}. \end{aligned} \tag{2.31}$$

To derive the average effects I first specify the conditional regressions for each group

$$\begin{aligned} E_{\text{Grp1}}(Y \mid Z) &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 + \beta_5 Z_1 + \beta_6 Z_2 \\ E_{\text{Grp2}}(Y \mid Z) &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_4 + \beta_7 Z_1 + \beta_8 Z_2 \\ E_{\text{Grp3}}(Y \mid Z) &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2. \end{aligned} \tag{2.32}$$

Comparing the first two groups against the reference group yields the two effect functions $g_{1-3}(Z) = E_{\text{Grp1}}(Y \mid Z) - E_{\text{Grp3}}(Y \mid Z)$ and $g_{2-3}(Z) = E_{\text{Grp2}}(Y \mid Z) - E_{\text{Grp3}}(Y \mid Z)$ as

$$\begin{aligned} g_{1-3}(Z) &= \beta_3 + \beta_5 Z_1 + \beta_6 Z_2 \text{ and} \\ g_{2-3}(Z) &= \beta_4 + \beta_7 Z_1 + \beta_8 Z_2. \end{aligned} \tag{2.33}$$

The two average effects of treatment one and treatment two (each compared to treatment three) are therefore

$$\begin{aligned} AE_{1-3} &= E\left(g_{1-3}(Z)\right) = \beta_3 + \beta_5\, E(Z_1) + \beta_6\, E(Z_2) \text{ and} \\ AE_{2-3} &= E\left(g_{2-3}(Z)\right) = \beta_4 + \beta_7\, E(Z_1) + \beta_8\, E(Z_2). \end{aligned} \tag{2.34}$$

Based on Equation 2.34, it is easy to formulate hypotheses about the average effects. For example, the null hypothesis H_0 stating that no (overall) average treatment effect is present may be written as: $H_0 :\ AE_{1-3} = AE_{2-3} = 0$, or more specifically

$$\begin{aligned} H_0 :\ \beta_3 + \beta_5\, E(Z_1) + \beta_6\, E(Z_2) = 0 \text{ and} \\ \beta_4 + \beta_7\, E(Z_1) + \beta_8\, E(Z_2) = 0. \end{aligned} \tag{2.35}$$

The H_0 can be tested with the linear hypothesis approach of the general linear model (multiple linear regression) if $E(Z_1)$ and $E(Z_2)$ are known. If they have to be estimated, the maximum likelihood method should be used. The advantage of the outlined procedures over the different ways of partitioning the sums of squares (the non-orthogonal ANOVA approach) is apparent: a hypothesis (about an average effect) does not change depending on the cell frequencies. Furthermore, H_0 is equivalent to the null hypothesis that is tested by ANOVA given an orthogonal design. Testing the main (treatment) effect in orthogonal ANOVA means testing the average (treatment) effect.

If randomization is successfully applied, the resulting average treatment effect is an unbiased estimate of the average causal effect in the sense of Rubin. This holds whether the design is orthogonal or non-orthorgonal. If randomization is not successfully applied then the outlined method tries to statistically adjust for bias with regard to estimating the average causal effect. However one can never be certain to achieve correct adjustment because there might be confounding variables omitted in the analysis (see the literature on causality, e. g., Steyer et al., 2007; Rosenbaum, 2002).

Again, the question whether multiple linear regression can be used to test the null hypothesis or whether the maximum likelihood procedure has to be applied, depends on whether the entries $E(Z_1) = P(Z = \text{level one})$ and $E(Z_2) = P(Z = \text{level two})$ are known or estimated. As mentioned before, these means might be considered to be known in some cases; if the covariate is under control by the researcher, if the whole population is included in the study, or if the parameters are inferred from some census data, or from previous surveys that used quota samples trying to perfectly reproduce the underlying population. Given that the means are (considered to be) known, H_0 is linear in the model parameters and multiple linear regression can be

used to test the hypothesis.

However, the covariate means $E(Z_1)$ and $E(Z_2)$ will often have to be estimated from the sample data and therefore the maximum likelihood method has to be applied. The Mplus input for the example is provided in Listing A.1 on page 141. Please note again that only the input for the restricted model is provided. The unrestricted model, just like in the example before, is equivalent to the saturated model tested by Mplus and therefore yields a chi-square value of zero. Hence, the chi-square value and the p value of the restricted model directly yield the test statistic and the p value corresponding to the null hypothesis.

2.6 Centering

In order to estimate and test average effects, centering (Marquardt, 1980) is a method that was proposed for designs with continuous covariates (see, e. g., Aiken & West, 1991). As we have seen in the first example, centering can help to facilitate the analysis of average effects. However, centering may sometimes lead to additional assumptions that are (simultaneously) tested. This will be shown with an example adapted from West et al. (1996).

This example includes a design with three treatment groups ($X = 0$, 1, 2) and two continuous covariates Z_1 and Z_2 (e. g., a pretest and a personality variable). In order to analyze the effects of the treatments on the dependent variable Y, the outcomes of treatment one and treatment two are compared to the outcomes of the control group ($X = 0$). Two dummy variables are used to code the treatment groups: $I_{X=1} = 1$, if treatment 1 and 0 else; $I_{X=2} = 1$, if treatment 2, and 0 else. The regression equation is structured

as

$$E(Y \mid X, Z_1, Z_2) = g_0(Z_1, Z_2) + g_1(Z_1, Z_2)I_{X=1} + g_2(Z_1, Z_2)I_{X=2}, \quad (2.36)$$

with the three functions g_0, g_{1-0} and g_{2-0}:

$$g_0(Z_1, Z_2) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 \quad (2.37)$$

$$g_{1-0}(Z_1, Z_2) = \beta_4 + \beta_5 Z_1 + \beta_6 Z_2 + \beta_7 Z_1 Z_2 \quad (2.38)$$

$$g_{2-0}(Z_1, Z_2) = \beta_8 + \beta_9 Z_1 + \beta_{10} Z_2 + \beta_{11} Z_1 Z_2. \quad (2.39)$$

This results in a regression with second order interactions

$$\begin{aligned} E(Y \mid X, Z_1, Z_2) &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 \\ &+\beta_4 I_{X=1} + \beta_5 Z_1 I_{X=1} + \beta_6 Z_2 I_{X=1} + \beta_7 Z_1 Z_2 I_{X=1} \\ &+\beta_8 I_{X=2} + \beta_9 Z_1 I_{X=2} + \beta_{10} Z_2 I_{X=2} + \beta_{11} Z_1 Z_2 I_{X=2}. \end{aligned} \quad (2.40)$$

The computation of the two average effects is straightforward. The average effect of treatment 1 vs. 0 and the average effect of treatment 2 vs. 0 are respectively

$$\begin{aligned} E\left(g_{1-0}(Z_1, Z_2)\right) &= \beta_4 + \beta_5\, E(Z_1) + \beta_6\, E(Z_2) + \beta_7\, E(Z_1 Z_2) \\ E\left(g_{2-0}(Z_1, Z_2)\right) &= \beta_8 + \beta_9\, E(Z_1) + \beta_{10}\, E(Z_2) + \beta_{11}\, E(Z_1 Z_2). \end{aligned}$$

Because $E(Z_1 Z_2) = Cov(Z_1, Z_2) + E(Z_1)\, E(Z_2)$, the null hypothesis stating that both average treatment effects are zero can be written as

$$\begin{aligned} H_0: \; 0 &= \beta_4 + \beta_5\, E(Z_1) + \beta_6\, E(Z_2) \\ &\quad +\beta_7 \left[Cov(Z_1, Z_2) + E(Z_1)\, E(Z_2)\right] \quad \text{and} \\ 0 &= \beta_8 + \beta_9\, E(Z_1) + \beta_{10}\, E(Z_2) \\ &\quad +\beta_{11} \left[Cov(Z_1, Z_2) + E(Z_1)\, E(Z_2)\right]. \end{aligned} \quad (2.41)$$

If the means $E(Z_1)$, $E(Z_2)$, and $Cov(Z_1, Z_2)$ are known, then the linear hypothesis test can be used by specifying a hypothesis matrix using these

means in order to test the null hypothesis. Otherwise, the outlined maximum likelihood procedure is applicable.

To see if centering facilitates the test of the null hypothesis, consider the centered[6] covariates $Z_1' = Z_1 - E(Z_1)$ and $Z_2' = Z_2 - E(Z_2)$. The equivalent regression model to Equation 2.6 is

$$\begin{aligned} E(Y \mid X, Z_1', Z_2') &= \gamma_0 + \gamma_1 Z_1' + \gamma_2 Z_2' + \gamma_3 Z_1' Z_2' \\ &+ \gamma_4 I_{X=1} + \gamma_5 Z_1' I_{X=1} + \gamma_6 Z_2' I_{X=1} + \gamma_7 Z_1' Z_2' I_{X=1} \\ &+ \gamma_8 I_{X=2} + \gamma_9 Z_1' I_{X=2} + \gamma_{10} Z_2' I_{X=2} + \gamma_{11} Z_1' Z_2' I_{X=2}. \end{aligned} \tag{2.42}$$

West et al. (1996) recommended to test the null hypothesis by conducting an R^2-difference test, which compares the full regression of Equation 2.6 to the restricted model

$$\begin{aligned} E_{res}(Y \mid X, Z_1', Z_2') &= \gamma_0 + \gamma_1 Z_1' + \gamma_2 Z_2' + \gamma_3 Z_1' Z_2' \\ &+ \gamma_5 Z_1' I_{X=1} + \gamma_6 Z_2' I_{X=1} + \gamma_7 Z_1' Z_2' I_{X=1} \\ &+ \gamma_9 Z_1' I_{X=2} + \gamma_{10} Z_2' I_{X=2} + \gamma_{11} Z_1' Z_2' I_{X=2}. \end{aligned} \tag{2.43}$$

Dropping the two terms $\gamma_4 I_{X=1}$ and $\gamma_8 I_{X=2}$ this R^2-difference tests the hypothesis

$$H_0' : \ \gamma_4 = \gamma_8 = 0. \tag{2.44}$$

To see if H_0' is equivalent to the H_0, consider the meaning of the regression coefficients γ_4 and γ_8. Substituting $Z_1 = Z_1' + E(Z_1)$ and $Z_2 = Z_2' + E(Z_2)$ in Equation 2.6 and applying some algebra, results in

$$\begin{aligned} \gamma_4 &= \beta_4 + \beta_5\, E(Z_1) + \beta_6\, E(Z_2) + \beta_7\, E(Z_1)\, E(Z_2) \text{ and} \\ \gamma_8 &= \beta_8 + \beta_9\, E(Z_1) + \beta_{10}\, E(Z_2) + \beta_{11}\, E(Z_1)\, E(Z_2). \end{aligned} \tag{2.45}$$

This shows that the hypothesis H_0' is only equivalent to the null hypothesis H_0 if $Cov(Z_1, Z_2) = 0$.

[6]Note that the centering is based the population means $E(Z_1)$ and $E(Z_2)$.

Given that there is no covariation between the covariates, the null hypothesis H_0 can be tested by a R^2-difference test comparing the model without the terms $I_{X=1}$ and $I_{X=2}$ to the full model. However, if a covariation between the two covariates is present, then the R^2-difference test does not test the H_0. This hypothesis can be tested using multiple linear regression with a hypothesis matrix that has $Cov(Z_1, Z_2)$ at the appropriate places.

If the means $E(Z_1)$, $E(Z_2)$, and the covariance $Cov(Z_1, Z_2)$ have to be estimated, and the centering is based on the sample means rather than the population mean, then the maximum likelihood procedure has to be applied. The corresponding Mplus input is given in Listing A.2 on page 142. The input is the same whether the covariates Z_1 and Z_2 are centered or not. The covariate means are treated as model parameters in both cases.

This example was given to emphasize that the linear hypothesis is more flexible than the R^2-difference test based on the centering approach, which is only applicable if there is no covariation between the covariates in the population. The linear hypothesis test however would still yield inflated Type-I errors if the covariate means and covariance are estimated. For this case the maximum likelihood approach is applicable.

2.7 Simulation

It was mentioned before that multiple linear regression methods might have drawbacks with regard to testing hypotheses about average treatment effects if interaction effects with the covariates are present, and if the means (and higher order moments) of the covariate that appear in the hypotheses have to be estimated.

A Monte Carlo Study is conducted in order to compare the performance

of multiple linear regression methods to the proposed maximum likelihood approach. A setting is chosen even simpler than the one of the introductory example. The data are generated to simulate a study that investigates the effect of a treatment on a continuous outcome variable Y with a randomized and balanced treatment-control group design. The treatment variable X indicates with one or zero, whether an observational unit is assigned to the treatment or the control group respectively. A continuous covariate is included in the study to investigate if the effect of the treatment depends on Z.

The data for the covariate is generated from a normal distribution with $E(Z) = 0$ and $Var(Z) = 4$. A sample size of 200 is chosen to resemble a typical sample size in applied psychological research. For the treatment variable X a random permutation of 100 zeros and 100 ones is drawn in order to implement random assignment. The data for the outcome variable $Y = E(Y \mid X, Z) + \epsilon$ is generated based on the following linear regression equation:

$$E(Y \mid X, Z) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX. \tag{2.46}$$

The values for the error variable ϵ are generated from a normal distribution with $E(\epsilon) = 0$ and $Var(\epsilon) = 16$. The mean $E(Z) = 0$ and the regression coefficients $\beta_0 = 0.5$, $\beta_1 = 1.8$, and $\beta_2 = 0$ are chosen so that the average effect of the treatment remains zero:

$$AE_{1-0} = E\left(g_{1-0}(Z)\right) = \beta_2 + E(Z)\beta_3 = 0. \tag{2.47}$$

This is important because the interaction effect β_3 varies between the values 0, 0.5, 1, 2.5, 4, 5, and 10. For each interaction effect value, 1000 data sets, also called replications, are generated. The average treatment effect is estimated and tested against zero with the following three methods:

1. *Centering on $E(Z)$, the Population Mean*: The values for $Z' = Z - E(Z)$, the covariate centered on the population mean, are computed. As described in the first example, the test is performed by estimating and testing the regression coefficient $\beta'_2 = \beta_2 + \beta_3\, E(Z)$ in the centered regression model $E(Y \mid X, Z') = \beta'_0 + \beta'_1 Z' + \beta'_2 X + \beta'_3 Z' X$ against zero using ordinary multiple regression (i. e. the general linear model).

2. *Centering on $\bar{Z}$, the Sample Mean*: The same procedure as in the previous method is applied to estimate and test the average treatment effect with the only difference that the centering of the covariate is based on the sample mean $\bar{Z}$ of each data set.

3. *Maximum Likelihood* (ML): The chi-square difference test is performed as described in the outline of the maximum likelihood test; comparing the unconstrained model with the model that has a non-linear constraint. The Mplus input file for the constrained model is given in Listing 2.1 on page 29. The estimated values for the average effects $E\left(g_{1-0}(Z)\right) = \beta_2 + \beta_3\, E(Z)$ are computed from the parameter estimates of the unconstrained model. The Mplus input file for the unconstrained model is equivalent to the input file of the constrained model without the constraint. The standard errors are computed using the multivariate delta method.

The performance of the three methods with regard to analyzing the average treatment effect is assessed with the following dependent measures. The average of the estimates across replications is computed in order to assess estimation bias. Standard error bias is assessed based on the average of the standard error estimators as well as the standard deviation of average effect estimators. The latter value is considered to be the population standard error

because the number of replications is 1000. Standard error bias is estimated by subtracting the population standard error value from the average standard error value and dividing this number by the population standard error value and multiplying by 100. An absolute value larger than 5 is typically considered to be biased (L. Muthén & Muthén, 2004). Finally the proportions of significant results at a significance level of $\alpha = 0.05$ are recorded in order to show whether the nominal alpha level is met or not.

Table 2.1 on page 41 shows the dependent measures as a function of the interaction effect. To facilitate the interpretation of the results an effect size measure for the interaction effect is provided. The effect size is computed dividing the interaction effect by the population standard deviation of the outcome variable given control. This standard deviation serves as a constant norm that is independent from the interaction effect, providing a measure or effect size that is in line with Cohen (1988).

The average estimates indicate that each method yields an unbiased estimate of the average treatment effect. The average standard error of the maximum likelihood method increases the larger the interaction effect. This is not the case for the two multiple linear regression methods. The only method yielding biased standard error estimators is the multiple regression method using the sample mean of the covariate. Given a small interaction the standard error bias seems inflated, however it is still below the cut point (-5). For medium and larger interaction effects the standard error estimators are severely biased; the bias increasing with the interaction effect. As a consequence of the biased standard errors, the proportion of significant results fail to meet the nominal alpha level. Instead they increase depending on the interaction effect.

Table 2.1

Dependent Measures of the Statistical Methods as a Function of the Interaction Effect

	Interaction effect						
	0	0.5	1	2.5	4	5	10
	Interaction effect size						
Method	0.00	0.09	0.19	0.46	0.74	0.93	1.86
	Average estimates						
$E(Z)$	0.01	0.00	0.02	0.01	0.00	−0.01	−0.02
$\bar{Z}$	0.01	−0.00	0.02	0.01	−0.02	−0.03	−0.03
ML	0.01	−0.00	0.02	0.01	−0.02	−0.03	−0.03
	Average standard errors						
$E(Z)$	0.57	0.57	0.57	0.57	0.57	0.57	0.57
$\bar{Z}$	0.57	0.57	0.57	0.57	0.57	0.57	0.57
ML	0.57	0.57	0.58	0.67	0.80	0.90	1.51
	Standard error bias						
$E(Z)$	−1.90	−0.04	−0.61	1.71	3.48	−1.92	0.30
$\bar{Z}$	−1.61	−0.52	−4.01	**−14.14**	**−26.73**	**−37.63**	**−63.09**
ML	−1.84	0.12	−1.22	1.16	3.61	−0.65	−1.33
	Proportion of significant tests						
$E(Z)$	0.05	0.06	0.06	0.04	0.04	0.06	0.05
$\bar{Z}$	0.05	0.06	0.06	**0.10**	**0.15**	**0.23**	**0.47**
ML	0.05	0.06	0.06	0.06	0.04	0.05	0.04

Note. Replications = 1000; $N = 200$. **Bold**: Standard error bias values below the nominal value −5 as well as proportions of significant tests differing significantly from the expected value of .05 (at a confidence level of .95 based on the exact binomial test).

2.8 Empirical Example: Effects of Insulation on Gas Consumption

A simple example will be used to illustrate the main ideas of this chapter. In the 1960's Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons. The first heating season was 26 weeks before cavity-wall insulation was installed, and the second heating season was 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption.

The data set is reported by Hand, Daly, Lunn, McConway, and Ostrowski (1994) and is available from the statistical programing environment R (R Development Core Team, 2006)[7]. The data set has 56 observations on the following three variables:

- Insul: A dichotomous variable, indicating whether the data were recorded before or after insulation. This variable corresponds to the treatment variable in the framework here.
- Temp: Purportedly[8] the average outside temperature in degrees Celsius. This variable corresponds to the covariate in the framework here.
- Gas: The weekly gas consumption in 1000s of cubic feet. This variable corresponds to the outcome variable in the framework here.

[7]The data frame is included in the `MASS` package under the name `whiteside`.

[8]It is noted in the description of the data set of the statistical programing environment R that the values are too low for any 56-week period in the 1960s in South-East England. It might be the weekly average of daily minima.

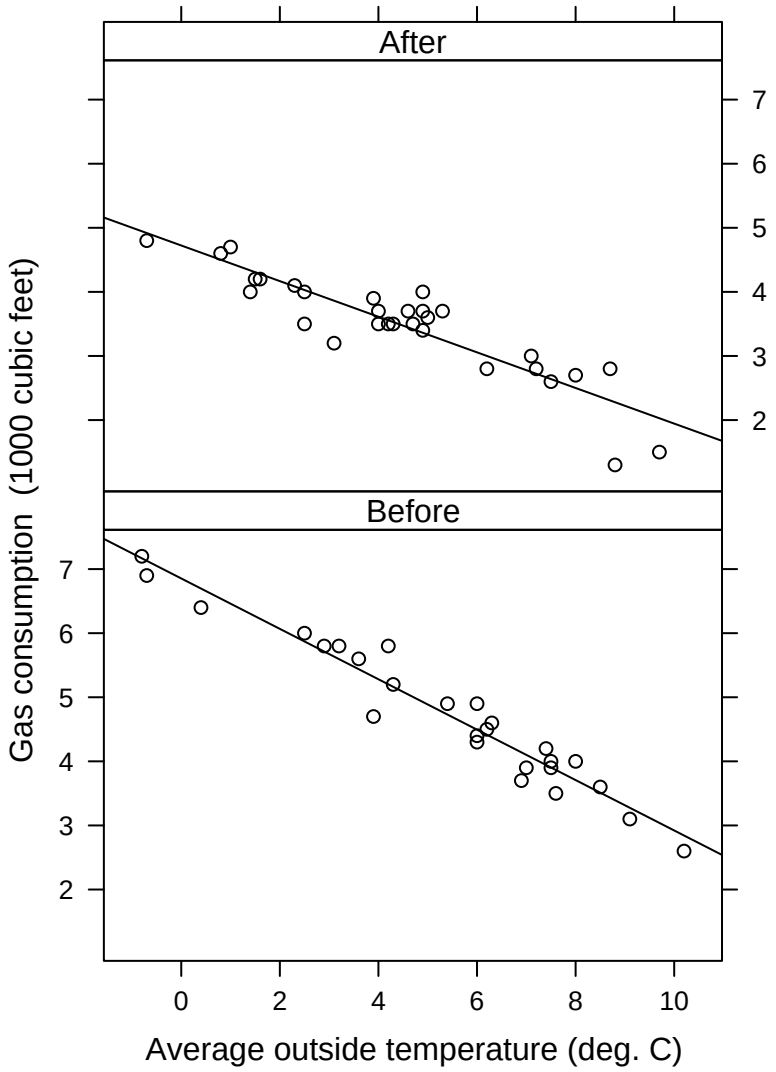


Figure 2.1: Gas consumption as a function of outside temperature before and after insulation was installed.

Table 2.2

Summary of the Regression Analysis Including the Centered Outside Temperature.

	$\hat{\beta}$	SE	t
Intercept	4.937^{**}	0.064	76.848
$Temp_c$	-0.393^{**}	0.022	-17.487
I_{After}	-1.568^{**}	0.088	-17.875
$I_{After}Temp_c$	0.115^{**}	0.032	3.591

Note. $R^2 = .93$. $^{*}p < .05$. $^{**}p < .01$

In Figure 2.1 the gas consumption is plotted against the outside temperature before and after the insulation. The figure also shows the estimated linear regression functions before and after the insulation, which appear to fit the data well. The relationship between gas consumption and outside temperature seems to be linear both before and after the treatment (for the considered range of temperature values). The slope of both regression lines is different before and after the insulation indicating an interaction between insulation and temperature.

As described previously, the covariate (here the outside temperature) is centered in order to facilitate the interpretation of the subsequent multiple regression analysis. The sample mean of the outside temperature (before and after the insulation) is $\overline{Temp} = 4.875$. Hence, the centered covariate is

$$Temp_C = Temp - \overline{Temp}. \tag{2.48}$$

Using ordinary least squares estimation the following regression equation is fitted to the data

$$E(Gas \,|\, Insul, Temp) = \beta_o + \beta_1 Temp_C + \beta_2 I_{After} + \beta_3 I_{After} Temp_C. \tag{2.49}$$

The results are given in Table 2.2. The overall model fit is very good with $R^2 = .93$. The estimated regression coefficient of the interaction term is 0.12 and significant at the .1 level. The effect size of the interaction is computed by taking the estimated regression coefficient of the interaction divided by the variance of the gas consumption before insulation. The value is 0.1 which is considered a small effect following Cohen's (1988) classification. The interaction indicates that the observed reduction in gas consumption after insulation is larger the colder the outside temperature becomes.

The estimated regression coefficient $\beta_2 = -1.57$ can be interpreted as the predicted difference between the gas consumption after insulation and the gas consumption before insulation for the mean outside temperature of 4.875. The corresponding hypothesis test indicates a rejection of the null hypothesis stating that the difference in gas consumption at the sample mean of the outside temperature is zero.

A second regression analysis based on the regression in Equation 2.49 but with the un-centered outside temperature was conducted. The following linear hypothesis

$$\begin{pmatrix} 0 & 0 & 1 & \overline{Temp} \end{pmatrix} \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \end{pmatrix}' = 0. \tag{2.50}$$

was applied. The tested hypothesis is equivalent to the previous hypothesis stating that the difference in gas consumption at the observed mean temperature is zero[9]. The results are exactly the same as in the previous test based on the centering method. The estimated value for this difference is about

[9]This is because

$$\begin{aligned} E(Gas \,|\, Insul = After, Temp = \overline{Temp}) - E(Gas \,|\, Insul = Before, Temp = \overline{Temp}) \\ = \beta_2 + \beta_3 \overline{Temp}. \end{aligned}$$

```
DATA: FILE IS whiteside.dat;
VARIABLE: NAMES ARE INSUL TEMP GAS;
USEVARIABLES ARE TEMP GAS IAFTER INT;
DEFINE: IAFTER = 0; IF(INSUL EQ 2) THEN IAFTER = 1;
DEFINE: INT = IAFTER * TEMP;
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
GAS ON TEMP IAFTER INT;
IAFTER WITH TEMP;
INT WITH TEMP;
INT WITH IAFTER;
OUTPUT: TECH1; TECH3;
SAVEDATA:
RESULTS ARE whitesideunc.dat;
TECH3 IS whitesideunct3.dat;
```

Listing 2.2: Mplus input for the Whiteside data.

-1.57 and the corresponding test statistic is about $F = 319.5$ ($df = 1$) which is equivalent to the squared t-value -17.87^2 of the centering method.

The maximum likelihood method is applied in order to analyze the average effect of the insulation and to compare the results with the centering method. The Mplus input is given in Listing 2.2 which estimates the following regression model

$$E(Gas \,|\, Insul, Temp) = \gamma_o + \gamma_1 Temp + \gamma_2 I_{After} + \gamma_3 I_{After} Temp. \quad (2.51)$$

The results are given in Table 2.3. The estimated interaction effect is identical to the previous results from the centering method. The corresponding standard error and t-value are almost identical. Applying the concept of the average treatment effect AE, the average effect (or the adjusted mean difference) of the insulation on gas consumption (with regard to the outside

Table 2.3

Summary of the Maximum Likelihood Analysis.

	Estimate	SE	t
γ_0	6.854^{**}	0.131	52.312
γ_1	-0.393^{**}	0.022	-18.147
γ_2	-2.130^{**}	0.174	-12.274
γ_3	0.115^{**}	0.031	3.726
$\overline{Temp}$	4.875^{**}	0.364	13.388
AE	-1.568^{**}	0.094	-16.613

Note. This model is saturated.

temperature), is specified as

$$AE = \gamma_2 + \gamma_3 \, E(Temp). \tag{2.52}$$

Using the parameter estimates from the ML method yields a value of about -1.568. The corresponding standard error is about 0.094. It was computed with the multivariate delta method (Cox & Hinkley, 1974; Stuard & Ord, 1994; Wasserman, 2004) using the estimated variance-covariance matrix of the model parameters from the ML estimation (requested in line 15 of Listing 2.2) and the gradient with respect to Equation 2.52. The corresponding t-value is -16.613 which is significant at the .01 level. Hence, the hypothesis of no average effect of the insulation can be rejected.

Comparing the results of the centering method with the ML method reveals that both methods provide similar results. The interaction and the corresponding standard error are estimated almost identical in both methods. The only notable difference occurs by comparing the results for the average effect from the ML estimation with the results for the β_2 regression

coefficient of the centering method. This comparison is of interest because estimating and testing β_2 based on the centering approach is sometimes used and recommended in order to estimate and test average treatment effects (here the average effect of the insulation). This was described in detail in section 2.1.

Both methods provide exactly the same estimate (-1.568). However, they differ with regard to the corresponding standard error. The standard error of the maximum likelihood method is 0.094 whereas the standard error of the centering method is 0.088. The standard error of the centering method is about 7.1% smaller than the standard error of the maximum likelihood method.

Based on the results of the previous discussions and the simulation study this outcome is as expected. The centering approach provides an unbiased estimate for β_2 which can be interpreted as the estimated conditional effect of the insulation at the mean of the outside temperature and is exactly the same as the estimated average effect of the ML method. The standard error of β_2 of the centering approach can be used to test β_2. However, because it is about 7% smaller than the standard error of the average effect from the ML estimation it should not be used to test the average effect of the insulation. Note that the effect size of the interaction is considered small. The results of the simulation study suggest that the difference in standard errors for the two compared methods would have been even larger for an interaction with a larger effect size.

The average effect estimated and tested by the ML method can be regarded as the adjusted mean difference of gas consumption before and after the insulation was installed. The adjustment is done by statistically controlling for temperature differences between treatment conditions. The (unad-

justed) mean difference of gas consumption after and before the insulation is about -1.267 which is about 19% smaller than the mean adjusted for the outside temperature. This result implies that the outside temperature is a confounding variable and should be taken into account when evaluating the effects of insulation on gas consumption. Under the assumption that there is no other confounding variable (besides the outside temperature) the average effect of the insulation (in other words the adjusted mean) can be interpreted as the average *causal* effect of the insulation on gas consumption.

2.9 Summary

We have seen how to specify average treatment effects in regression models that include a treatment-covariate interaction with regard to an outcome variable. It was shown that existing methods that center the covariate provide a method to estimate and test this average treatment effect under certain conditions. It was also argued that the linear hypothesis test of the general linear model also provides this test and is applicable to more cases. An example was given for a regression model that involves higher order interaction effects. It was also argued that centering as well as the linear hypothesis yield inflated Type I errors when the covariate means are estimated from the sample. Both methods treat the covariate means as fixed terms and not as model parameters.

The outlined maximum likelihood procedure is an important contribution to the methodology of the analysis of treatment effects, because in applied settings it will often be the case that the mean $E(Z)$ of the covariate will have to be estimated from the sample. The simulation shows that the maximum likelihood procedure provides unbiased estimates of the average treatment

effects as well as its standard error. The maximum likelihood procedure should be given advantage over the multiple regression method especially if medium or large interaction effects are present.

In summary the outlined procedure allows to incorporate covariates in the analysis of treatment effects. It allows to study how the treatment effect depends on covariates and also to estimate and test the average effect of a treatment. The main limitation of the approach is that measurement errors of the covariates are not accounted for. Many covariates in the social sciences are however only measurable with a measurement error and it is well known that such fallible covariates may bias the analysis of treatment effects (see, e. g. Bollen, 1989; Cohen et al., 2003). The solution to this measurement problem is to use structural equation modeling where the covariates are latent variables. The following chapters introduce models that allow to estimate and test average treatment effects if interaction effects between treatment and latent covariates are present.

Chapter 3

Average Effects in Latent Variable Modeling

3.1 A Latent Variable Model Involving Interaction between Treatment and Latent Covariate

This section describes how the concept of average treatment effects can be applied for latent variable models that involve interactions between treatments and latent covariates. The latent variable models considered include continuous outcome variables. It is clear however, that methodology for categorical or ordinal variables is needed in practice as well. If the outcome is simply a continuous manifest variable, the procedures outlined here are applicable simply by replacing the latent outcome with the manifest outcome.

Consider a study that is designed to investigate the effects of a treatment on a latent outcome variable η measured by two observed variables Y_1 and Y_2. The study is based on a treatment-control-group design. The two groups

are represented by the treatment variable X with values $j = 1, 2$. The study includes a continuous latent covariate ξ measured by two continuous observed variables Z_1 and Z_2. The continuous latent covariate may represent a latent pretest variable that is measured by two pretest measures before the treatment is applied. This design is frequently used in the social science research and has therefore been called *workhorse design* (see, e. g., Shadish et al., 2002).

The interaction between the treatment and the latent covariate is expressed in terms of the regression equation

$$E(\eta \,|\, X, \xi) = \alpha + \beta_1 I_{X=2} + \beta_2 \xi + \beta_3 I_{X=2} \xi. \tag{3.1}$$

The residual is defined in the usual way as $\zeta := \eta - E(\eta \,|\, X, \xi)$. $I_{X=2}$ is an indicator variable defined as

$$I_{X=2} := \begin{cases} 1 & \text{if } X = 2, \\ 0 & \text{else,} \end{cases} \tag{3.2}$$

so that group one serves as the control group. Throughout the following discussion, group one will be referred to as the control (or reference) group and group two as the treatment group.

The latent covariate ξ is measured by two observed variables Z_1 and Z_2 with the following measurement model[1]

$$\begin{aligned} Z_1 &= \nu_{Z_1} + \lambda_{Z_1} \xi + \epsilon_{Z_1}, \\ Z_2 &= \nu_{Z_2} + \lambda_{Z_2} \xi + \epsilon_{Z_2}, \\ Cov(\xi, \epsilon_{Z_1}) &= Cov(\xi, \epsilon_{Z_2}) = Cov(\epsilon_{Z_1}, \epsilon_{Z_2}) = 0. \end{aligned} \tag{3.3}$$

[1] A common way to identify this measurement model is to set $\nu_{Z_1} = 0$ and $\lambda_{Z_1} = 1$. Identification issues of the measurement models are discussed later.

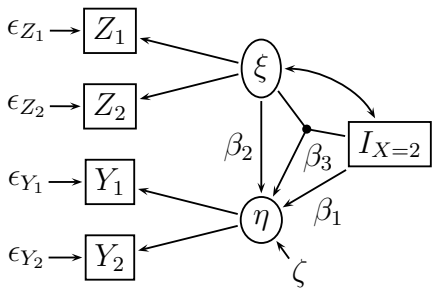


Figure 3.1: Latent variable model involving an interaction between a dichotomous treatment variable and a continuous latent covariate.

The two observed variables Y_1 and Y_2 measure the latent outcome η with the following measurement model:

$$\begin{aligned} Y_1 &= \nu_{Y_1} + \lambda_{Y_1}\eta + \epsilon_{Y_1}, \\ Y_2 &= \nu_{Y_2} + \lambda_{Y_2}\eta + \epsilon_{Y_2}, \\ Cov(\eta, \epsilon_{Y_1}) &= Cov(\eta, \epsilon_{Y_2}) = Cov(\epsilon_{Y_1}, \epsilon_{Y_2}) = 0. \end{aligned} \tag{3.4}$$

Both measurement models represent the model of τ-congeneric variables (for a more detailed description of these measurement models see Steyer, 2001 or Steyer & Eid, 2001). The measurement residuals ϵ_{Z_1}, ϵ_{Z_2}, ϵ_{Y_1}, and ϵ_{Y_2} are assumed to be uncorrelated.

Except for the interaction of the latent covariate with the treatment, this is a standard SEM model. It is represented by the path diagram in Figure 3.1. The arrow labeled with β_3 which begins at ξ and $I_{X=2}$ and points to η describes the interaction. Another way to represent the interaction is given in Figure 3.2. The arrow with the dotted line emphasizes that the effect of $I_{X=2}$ depends on ξ. The corresponding label of the arrow gives the effect function instead of the regression coefficient of the interaction.

This latent variable model may, for example, describe the interaction between a treatment and a pretest with regard to post-test, where both the

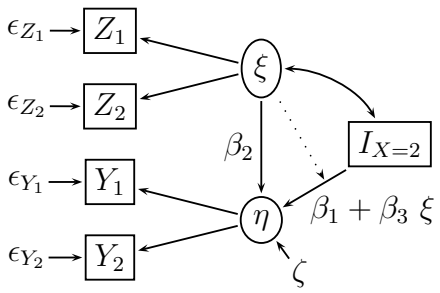


Figure 3.2: An alternative way to represent a latent variable model involving an interaction between a dichotomous treatment variable and a continuous latent covariate.

pretest and the post-test are variables that are measured with a measurement error. The regression in Equation 3.1 is similar to an ordinary multiple regression with an interaction term. Most readers will be familiar with this multiple regression approach involving manifest variables, often termed *moderated multiple regression* (Aiken & West, 1991; Jaccard et al., 1990; Judd & McClelland, 1989; Saunders, 1956). Like in the manifest regression the interpretation of Equation 3.1 is as follows. With an interaction effect present, β_1, the *first-order effect* of the treatment, should in general not be interpreted as the main effect or the average effect of the treatment.

To establish the meaning of β_1 consider the regressions of η on ξ for each group

$$\begin{aligned} E_{X=1}(\eta \,|\, \xi) &= \alpha + \beta_2 \xi \\ E_{X=2}(\eta \,|\, \xi) &= \alpha + \beta_1 + \beta_2 \xi + \beta_3 \xi. \end{aligned} \tag{3.5}$$

The difference of these two regressions for a given value of ξ is called *conditional effect* of the treatment. The function

$$g_{2-1} := E_{X=2}(\eta \,|\, \xi) - E_{X=1}(\eta \,|\, \xi) = \beta_1 + \beta_3 \xi \tag{3.6}$$

is called *effect function*. It is apparent that β_1 represents the conditional effect

of the treatment at the value 0 of the latent covariate. Thus, the meaning of the latent covariate value 0 must be considered in order to interpret β_1. Hence, scaling ξ so that $E(\xi) = 0$ ensures that the interpretation of β_1, the first order-effect of the treatment, occurs at a meaningful value of the latent covariate.

To return to the concept of the average treatment effect, the question is, if β_1 can be interpreted as the average (or the main) effect of the treatment if $E(\xi) = 0$. To answer this question, a definition of the average effect of the treatment is required. Consider $g_{2-1}(\xi)$, the effect function (given in Equation 3.6), that maps each covariate value to the conditional treatment effect. It is clear that the expected treatment effect depends (linearly) on the latent covariate. The latent covariate is called a *moderator* (see, e. g., Baron & Kenny, 1986).

For the given example, AE_{2-1}, the average effect of treatment two vs. control is specified as the average of the conditional effect function

$$AE_{2-1} = E\left(g_{2-1}(\xi)\right) = E(\beta_1 + \beta_3\xi) = \beta_1 + \beta_3\, E(\xi). \tag{3.7}$$

The function g_{2-1} maps every value of ξ to the conditional treatment effect. This function is as a random variable. Hence, the average effect of the treatment is defined as the average (or the expected value) of this random variable. If no interaction is present (i. e. $\beta_3 = 0$), then the conditional treatment effects are the same across all values of ξ. Consequently the average treatment is equal to β_1, the constant value of the conditional treatment effects.

If an interaction is present however, Equation 3.7 shows that the average treatment effect is a (non-linear) function of the three parameters β_1, β_3, and $E(\xi)$. It is obvious that scaling ξ so that $E(\xi) = 0$ ensures that β_1, the first order effect of the treatment, is equal to the average treatment effect. This

will be important throughout the following chapters.

3.2 Generalizations

The example of the last section included only one (continuous) covariate and only two treatments (or treatment groups). The concept may be generalized to more than two treatments and to possibly multiple latent covariates.

Let there be J treatment groups. The treatment variable X may take on the value $X = 1$ for control, and $X = 2, \ldots, X = J$ for the treatment groups. The following $J - 1$ dummy codes may be used to identify each treatment: $I_{X=2}, \ldots, I_{X=J}$ indicate with 1 and 0 whether or not the observational unit is assigned to treatment j, with $j = 2, \ldots, J$.

The regression equation of the outcome variable Y regressed on the treatment and the latent covariates can generally be written as

$$E(\eta \,|\, X, \xi) = g_1(\xi) + g_{2-1}(\xi) \cdot I_{X=2} + \cdots + g_{J-1}(\xi) \cdot I_{X=J}, \tag{3.8}$$

where ξ represents a univariate or multivariate latent covariate. Consequently there are $J - 1$ average effects, which are defined as:

Definition 2. *The average effect AE_{j-1} of treatment j (compared to the control group $X = 1$) with regard to the covariate ξ is defined as the average of the effect function $g_{j-1}(\xi)$,*

$$AE_{j-1} = E\left(g_{j-1}(\xi)\right), \tag{3.9}$$

with $j = 2, \ldots, J$.

Note again that the average effect is defined with regard to the covariate ξ. For a different covariate the average treatment effect might differ. However, if the regression is causally unbiased (e. g. given a randomized design),

then it can be shown that the average treatment effect is the same for all (combinations) of possible covariates (see Steyer et al., 2007, for a detailed discussion of this topic on causality).

In the following, I describe methods to estimate and to test the average treatment effect for the two-group example. Two cases are treated separately. First, chapter 3.3 treats the case where randomization is successfully implemented in the study. Standard multiple group analysis is applicable by imposing certain constraints on the model parameters. Second, chapter 4 describes methods on how to analyze average treatment effects for non-experimental designs, in other words designs, where randomization is not (successfully) implemented.

3.3 Average Treatment Effects in Randomized Studies

3.3.1 A General Latent Variable Framework

The model represented by Equations 3.1 - 3.4, includes a (linear) interaction between a dichotomous observed variable $I_{X=2}$ and a continuous latent variable ξ. Models including this type of interaction are typically analyzed with the following general latent variable framework (cf. Bollen, 1989; Jöreskog & Sörbom, 1979; Sörbom, 1978). For treatment (or in general population) j, consider a p-dimensional observed variable vector $\boldsymbol{y}^{(j)}$ related to an m-dimensional latent variable vector $\boldsymbol{\eta}^{(j)}$ through a factor-analytic measurement model

$$\boldsymbol{y}^{(j)} = \boldsymbol{\nu}^{(j)} + \boldsymbol{\Lambda}^{(j)}\boldsymbol{\eta}^{(j)} + \boldsymbol{\epsilon}^{(j)}, \tag{3.10}$$

where $\boldsymbol{\nu}^{(j)}$ is a vector of measurement intercepts, $\boldsymbol{\Lambda}^{(j)}$ is a $p \times m$-dimensional matrix of measurement slopes (factor loadings), and $\boldsymbol{\epsilon}^{(j)}$ is a p-dimensional vector of measurement residuals. The variance-covariance matrix of the measurement residuals is $Var(\boldsymbol{\epsilon}^{(j)}) = \boldsymbol{\Theta}^{(j)}$.

The latent variables have the structural relations

$$\boldsymbol{\eta}^{(j)} = \boldsymbol{\alpha}^{(j)} + \mathbf{B}^{(j)}\boldsymbol{\eta}^{(j)} + \boldsymbol{\zeta}^{(j)}, \tag{3.11}$$

where $\boldsymbol{\alpha}^{(j)}$ is an m-dimensional vector of structural intercepts (for endogenous latent variables) or means (for exogenous latent variables), $\mathbf{B}^{(j)}$ is an m-dimensional matrix of structural slopes, and $\boldsymbol{\zeta}^{(j)}$ is and m-dimensional vector of structural residuals. $Var(\boldsymbol{\zeta})^{(j)} = \boldsymbol{\Psi}^{(j)}$ is a residual (for endogenous latent variables) or latent variable covariance matrix (for exogenous latent variables).

Under regular assumptions on the residuals, we have the mean and the covariance structure

$$E\left(\boldsymbol{y}^{(j)}\right) = \boldsymbol{\mu}^{(j)} = \boldsymbol{\nu}^{(j)} + \boldsymbol{\Lambda}^{(j)}\left(\boldsymbol{I} - \mathbf{B}^{(j)}\right)^{-1}\boldsymbol{\alpha}^{(j)} \tag{3.12}$$

and

$$\begin{aligned} Cov\left(\boldsymbol{y}^{(j)}\right) = \boldsymbol{\Sigma}^{(j)} = \\ = \boldsymbol{\Lambda}^{(j)}\left(\boldsymbol{I} - \mathbf{B}^{(j)}\right)^{-1}\boldsymbol{\Psi}^{(j)}\left(\boldsymbol{I} - \mathbf{B}^{(j)}\right)^{-1'}\boldsymbol{\Lambda}^{(j)'} + \boldsymbol{\Theta}^{(j)}. \end{aligned} \tag{3.13}$$

This is the standard multiple-group structural equation modeling framework. With the customary assumption of i.i.d. sampling from each of the J populations, a simultaneous, multiple-group (multiple-population) analysis is commonly achieved by minimizing the fitting function F

$$\begin{aligned} F = \\ \sum_{j=1}^{J}\Bigg(N^{(j)}\left[\ln|\boldsymbol{\Sigma}^{(j)}| + tr\left(\boldsymbol{\Sigma}^{(j)^{-1}}\boldsymbol{T}^{(j)}\right) - \ln|\boldsymbol{S}^{(j)}| - p\right]\Bigg)/(N-1), \end{aligned} \tag{3.14}$$

where N is the total sample size and

$$\boldsymbol{T}^{(j)} = \boldsymbol{S}^{(j)} + \left(\bar{\boldsymbol{y}}^{(j)} - \boldsymbol{\mu}^{(j)}\right)\left(\bar{\boldsymbol{y}}^{(j)} - \boldsymbol{\mu}^{(j)}\right)', \tag{3.15}$$

which gives maximum-likelihood estimation under multivariate normality for $\boldsymbol{y}^{(j)}$ (see, e. g., Jöreskog & Sörbom, 1979; Sörbom, 1982). At the optimal value of F, $(N-1)F$ has asymptotically a chi-squared distribution.

3.4 A Standard Multi-group Model

The (single-group) model represented by Equations 3.1 - 3.4 may be fitted into the general latent modeling framework (Sörbom, 1978). For $j = 1, 2$ consider the structural part

$$\eta^{(j)} = \alpha^{(j)} + \beta^{(j)}\xi^{(j)} + \zeta^{(j)}, \tag{3.16}$$

the measurement models[2]

$$\begin{aligned} Y_1^{(j)} &= \nu_{Y_1}^{(j)} + \lambda_{Y_1}^{(j)}\eta^{(j)} + \epsilon_{Y_1}^{(j)}, \\ Y_2^{(j)} &= \nu_{Y_2}^{(j)} + \lambda_{Y_2}^{(j)}\eta^{(j)} + \epsilon_{Y_2}^{(j)}, \\ Z_1^{(j)} &= \nu_{Z_1}^{(j)} + \lambda_{Z_1}^{(j)}\xi^{(j)} + \epsilon_{Z_1}^{(j)}, \\ Z_2^{(j)} &= \nu_{Z_2}^{(j)} + \lambda_{Z_2}^{(j)}\xi^{(j)} + \epsilon_{Z_2}^{(j)}, \end{aligned} \tag{3.17}$$

and the additional assumptions

$$\begin{aligned} Cov^{(j)}(\xi, \epsilon_{Z_1}) = Cov^{(j)}(\xi, \epsilon_{Z_2}) = Cov^{(j)}(\epsilon_{Z_2}, \epsilon_{Z_2}) = 0, \\ Cov^{(j)}(\eta, \epsilon_{Y_1}) = Cov^{(j)}(\eta, \epsilon_{Y_2}) = Cov^{(j)}(\epsilon_{Y_2}, \epsilon_{Y_2}) = 0, \\ Cov^{(j)}(\epsilon_{Z_1}, \epsilon_{Y_1}) = Cov^{(j)}(\epsilon_{Z_1}, \epsilon_{Y_2}) = Cov^{(j)}(\epsilon_{Z_2}, \epsilon_{Y_1}) = \\ = Cov^{(j)}(\epsilon_{Z_2}, \epsilon_{Y_2}) = 0. \end{aligned} \tag{3.18}$$

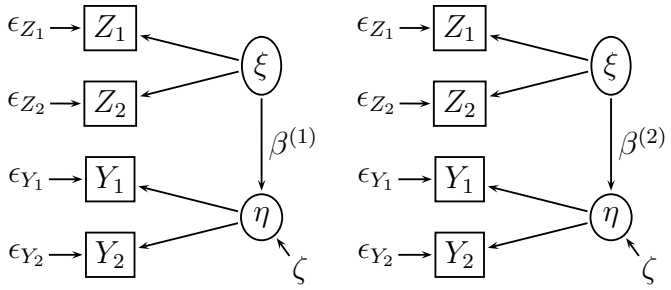
[2]The measurement models may be identified for example by setting $\nu_{Y_1}^{(1)} = \nu_{Z_1}^{(1)} = 0$ and $\lambda_{Y_1}^{(1)} = \lambda_{Z_1}^{(1)} = 1$. The identification of the measurement models is described later.

Let $\boldsymbol{y}$ in Equation 3.10 contain all observed variables Y_1, Y_2, Z_1, and Z_2 and $\boldsymbol{\eta}$ contain the two latent variable η and ξ. The parameters of Equations 3.10 and 3.11 are as follows: $\boldsymbol{\nu}^{(j)}$ contains the measurement intercepts $\nu_{Y_1}^{(j)}$, $\nu_{Y_2}^{(j)}$, $\nu_{Z_1}^{(j)}$, and $\nu_{Z_2}^{(j)}$ of each group; $\boldsymbol{\Lambda}^{(j)}$ contains 0s and the measurement slopes $\lambda_{Y_1}^{(j)}$, $\lambda_{Y_2}^{(j)}$, $\lambda_{Z_1}^{(j)}$, and $\lambda_{Z_2}^{(j)}$; $\boldsymbol{\Theta}^{(j)}$ contains the variance-covariance matrix of the measurement residuals $\epsilon_{Y_1}^{(j)}$, $\epsilon_{Y_2}^{(j)}$, $\epsilon_{Z_1}^{(j)}$, and $\epsilon_{Z_2}^{(j)}$; $\boldsymbol{\alpha}$ contains $\alpha^{(j)}$ the structural intercept of each group and $E^{(j)}(\xi)$, the mean of the covariate of each group; $\mathbf{B}$ contains 0s and the structural slopes $\beta^{(j)}$; finally $\boldsymbol{\Psi}$ contains $Var^{(j)}(\zeta)$ the variance of the structural residual as well as $Var^{(j)}(\xi)$ the variance of the latent covariate of each group.

This two-group model may be described in path diagram form as shown in Figure 3.3. It is straightforward how to generalize this two-group model to multiple groups. Although the single-group model (Equations 3.1 - 3.4) may be analyzed with the multi-group model in the outlined way, the two models are not identical. The multi-group model is more general in that it enables population (or group) differences with regard to the variances of the structural residual. These differences are not offered by the single-group model.

In the following, it is shown how the multi-group approach can be used to analyze treatment effects for randomized studies if interactions between treatment and latent covariates are present. For the sake of simplicity and continuity the examples are done with Mplus.

Consider an intervention study where individuals are measured before being randomized into a treatment or a control group and then measured thereafter. In line with Jöreskog and Sörbom (1979), this may be viewed as data from two different populations. The control group population represents the normative set of outcomes that would have been observed also



(a) Group 1 (b) Group 2

Figure 3.3: A multi-group model to analyze interaction between treatment and a latent covariate.

in the treatment group had the individuals not been chosen for treatment. The effect of treatment is assessed by comparing the outcomes in the treatment population with those in the control population. Treatment effects are assessed by comparing the regressions of η on ξ for each group:

$$\begin{aligned} E_{X=1}(\eta \,|\, \xi) &= \alpha^{(1)} + \beta^{(1)}\xi \\ E_{X=2}(\eta \,|\, \xi) &= \alpha^{(2)} + \beta^{(2)}\xi. \end{aligned} \tag{3.19}$$

The measurement models of the single-group approach and the multi-group approach are equivalent if the parameters of the multi-group approach are set invariant across the groups. Because it simplifies the identification of the multi-group models considerably and because the focus here is on the treatment effect rather than measurement equivalence, the measurement models of the groups are treated invariant throughout the discussion of the multi-group approach.

The structural equations of the multi-group model are equivalent to the

structural equation of the single-group model (see Equation 3.5), if

$$\begin{aligned} \alpha^{(1)} &= \alpha & \beta^{(1)} &= \beta_2 \\ \alpha^{(2)} &= \alpha + \beta_1 & \beta^{(2)} &= \beta_2 + \beta_3. \end{aligned} \tag{3.20}$$

The interpretation of the parameters is therefore identical to the single-group model. An interaction effect is present if $\beta^{(2)} - \beta^{(1)} \neq 0$. Given an interaction, $\alpha^{(2)} - \alpha^{(2)}$ can be interpreted as the treatment effect at the value 0 of ξ but in general not as the average treatment effect.

The effect function for the multi-group model is computed by the difference of the group specific structural regressions (see Equation 3.19)

$$\begin{aligned} g_{2-1}(\xi) &= E_{X=2}(\eta \,|\, \xi) - E_{X=1}(\eta \,|\, \xi) \\ &= \alpha^{(2)} - \alpha^{(1)} + \left(\beta^{(2)} - \beta^{(1)}\right) \xi. \end{aligned} \tag{3.21}$$

The average treatment effect is specified as the mean of the the effect function

$$AE_{2-1} = E[g_{2-1}(\xi)] = \alpha^{(2)} - \alpha^{(1)} + \left(\beta^{(2)} - \beta^{(1)}\right) E(\xi). \tag{3.22}$$

Usually hypotheses in SEM are about (single) parameters or linear combinations of these. For example, the hypothesis assuming that no interaction effect is present can be written as $\beta^{(2)} - \beta^{(1)} = 0$ (see, e. g., Jaccard & Wan, 1996). Equation 3.22 however shows that the average effect is a function of the two structural intercepts and the two structural slopes with the two structural slopes multiplied by the mean of the latent covariate. Whereas $\alpha^{(1)}$, $\alpha^{(2)}$, $\beta^{(1)}$, and $\beta^{(2)}$ are parameters of the multi-group model, $E(\xi)$, the (grand-) mean of the latent covariate, is not a parameter of the multi-group model. Without further adjustment this would make the multi-group model not feasible to analyze average treatment effects. I describe solutions that allow to apply the multi-group approach to the analysis of average effects.

The following section treats randomized designs. Chapter 4 will then focus on non-randomized designs and the comparison between multi-group and single-group approach.

Based on a randomized design it is possible to use the multi-group analysis to analyze average treatment effects. As a consequence of randomization the treatment assignment is independent from ξ. Hence the expected values of the latent covariate are equal across the two populations. It is therefore justified to set these means equal and use the corresponding estimate as an estimate for the grand mean of the latent covariate. It is also possible to set the variance of the latent covariate equivalent across the groups. This will however not be considered here because it is irrelevant for analyzing average treatment effects.

3.4.1 Scaling the Latent Variables

Given randomization, the key principle of the multi-group approach in order to analyze average treatment effects is to pose equality constraints on the group means of the latent covariate

$$E(\xi \mid X{=}1) = E(\xi \mid X{=}2). \tag{3.23}$$

For the sake of simplicity, the parameters of the measurement models of both latent variables are restricted to be equal across groups. For $k = 1, 2$ we have

$$\nu_{Z_k}^{(1)} = \nu_{Z_k}^{(2)} \qquad \lambda_{Z_k}^{(1)} = \lambda_{Z_k}^{(2)} \tag{3.24}$$

$$\nu_{Y_k}^{(1)} = \nu_{Y_k}^{(2)} \qquad \lambda_{Y_k}^{(1)} = \lambda_{Y_k}^{(2)}. \tag{3.25}$$

The latent variables are typically scaled by setting (at least) one measurement slope of each measurement model to one, for example $\lambda_{Y_1}^{(1)} = \lambda_{Z_1}^{(1)} = 1$. I will describe two ways to complete the scaling of the latent variables and the consequences for the analysis of the average effect.

Centering the Latent Covariate

One way to complete the scaling of the latent covariate ξ is to set the mean of the latent covariate in group one to zero

$$E(\xi \mid X{=}1) = 0. \tag{3.26}$$

Because the means of ξ are set equal across the groups (Equation 3.23), the overall mean of ξ is zero

$$E(\xi) = 0. \tag{3.27}$$

The scaling of the latent outcome variable may be completed by fixing the structural intercept in group one to zero

$$\alpha^{(1)} = 0. \tag{3.28}$$

The main advantage of this scaling is that the identification of the average effect (see Equation 3.22) is reduced to the single parameter

$$AE_{2-1} = \alpha^{(2)}, \tag{3.29}$$

the intercept of the latent outcome variable in the second group. Hence, the average treatment effect can be estimated and tested by estimating and testing the single parameter $\alpha^{(2)}$.

Listing A.3 on page 143 gives an example for a Mplus input. The corresponding Mplus output includes information about the overall model fit as well as estimates and tests for each parameter of the model, including standard errors and t-values. The parameters of special interest here are of course $\beta^{(1)}$ and $\beta^{(2)}$, the structural slopes the two groups. The difference between these two parameters is an estimate for the interaction effect. The interaction will be further discussed below.

If no interaction effect is present, the effect of the treatment is constant across all values of the latent covariate and equal to $\alpha^{(2)}$. It is trivial that the average treatment effect is equal to this constant measure.

The outline of the last paragraphs however has shown that the average treatment effect is also equal to $\alpha^{(2)}$, if an interaction effect is present. Besides the estimate and the standard error Mplus also prints the estimate divided by the standard error, a test statistic that is approximately normally distributed (z-score) in large samples, which may be used to test the null hypothesis stating that no average treatment effect is present. Testing the average treatment effect against values different than zero is also possible. The difference between the estimate of $\alpha^{(2)}$ and the value that it is tested against (assuming this value is the true population parameter) divided by the estimated standard error is a test statistic which is (approximately) normally distributed (z-score) in large samples.

Testing the interaction may be done by a chi-square difference test between the model specified in Listing A.3 on page 143 and the identical model with an additional equality constraint on the two structural slopes (Jaccard & Wan, 1996). Programs like Mplus (Version 4) and LISREL also offer the option to directly test the interaction, without the need of a second restricted model. These programs offer to specify an additional parameter that may be tested against a certain value. Listing A.4 on page 143 gives the corresponding Mplus input. Lines 1 to 8 are identical to the ones in Listing A.3. Lines 9 to 13 implement the additional test for the interaction.

The output of Listing A.4 is identical to the output of Listing A.3 except for the additional parameter specifying the interaction. The same information is provided for the interaction effect as it is provided for all model parameters, including the estimate and the standard error of the interaction

effect. And the estimate divided by the standard error is again a test statistic that is approximately normally distributed (z-score) in large samples. Testing the interaction effect against values different than zero is again done by subtracting the hypothetical true population value from the estimated interaction effect and dividing it by the standard error.

Using this method it is possible to simultaneously test the interaction effect and the average treatment effect. In this way the analysis reveals if and how the effect of the treatment on the outcome depends on the latent covariate. Given this interaction model Jaccard and Wan (1996) suggested to ignore the latent covariate in order to test the average effect. The authors though it would be impossible to analyse the average treatment effect (also referred to as main effect) with a model that includes the interaction. The outlined method shows however that it is possible to analyze the average treatment effect without ignoring the interaction with the latent covariate.

An Alternative Scaling Method

There are alternative ways to scale the latent variables. Instead of setting the mean of the latent covariate and the structural intercept of the latent outcome in group one to zero they may be set free and one measurement intercept of each measurement model may be set to zero, for example $\nu_{Y_1}^{(1)} = \nu_{Z_1}^{(1)} = 0$. Note that the group means of the latent covariate are still set equal across groups. Given this scaling, the average treatment effect can be estimated with the estimates of the involved parameters according to Equation 3.22. Testing the average treatment effect requires the specification of an additional parameter. The corresponding Mplus input given in Listing A.5 on page 144.

Listing A.5 is contains more statements than Listing A.3 because the alternative scaling method requires more parameter specifications that differ

from the default settings. The measurement intercepts are set free and equal across groups per default[3]. Line 7 and 8 of Listing A.5 specify the first measurement intercepts for both measurement models to zero in both groups. The measurement slopes are set equal across groups per default. The first measurement slopes of each measurement model, $\lambda_{Y_1}^{(1)}$ and $\lambda_{Z_1}^{(1)}$, are set to one per default. Hence, both inputs do not need further statements in order to specify tau-congeneric measurement models. The measurement errors of the observed variables are set equal across groups and are assumed to be uncorrelated with each other.

The group means of ξ in both inputs are set equal across groups. In Listing A.3 the group means of ξ are set to zero in line 7 with `[XI@0];` whereas in Listing A.5 they are constrained to be equal by placing the same name in parentheses following the parameter specification in each group (line 10 and 13). The structural intercepts and slopes in Listing A.5 are set free and are named by placing a name in parentheses after the parameter in order to use them for the additional (or new) parameter that specifies the average effect.

It is also important to note that the multi-group approach is more general than the single-group model given in Equations 3.16 to 3.17 because it allows several parameters to differ between the groups, that are implicitly treated constant in the single-group model. The variance of the structural residual, for example, may be different in the control group than in the treatment group.

Both inputs (i. e. models) yield the same model fit and the same results for the average treatment effect. As mentioned before, the test of the interaction effect may be included in the analysis by specifying an additional

[3]Hence, no statements are required in Listing A.3

(or new) parameter using the difference between the two structural slopes of the two groups. Testing the interaction may be included in the alternative scaling version by adding an additional parameter in the same way it is done in Listing A.4.

We have seen that the general latent modeling framework can be used to examine average (or main) treatment effects and interaction effects simultaneously. The key principle was to constrain the means of the latent covariate to be equal across groups, which is feasible in randomized designs. Traditionally, such an analysis was considered not possible (see, e. g., Jaccard & Wan, 1996, p. 41) and interaction effects were examined separately from main (or average) effects. In a randomized design it is possible to perform the test of the average treatment effect separately simply by ignoring the latent covariate and comparing the group means of the (latent or observed) outcome variables. Ignoring the latent covariate however may lead to an increase in residual variance and to a considerable loss in power. This will be discussed in the next section.

3.5 Power for Randomized Designs

For randomized designs, it is possible to analyze the average effect of the treatment simply by comparing the outcome means of each group. Ignoring the latent covariate in the analysis yields an causally unbiased estimation of the average treatment effect (see Steyer et al., 2007, for details). However, including the latent covariate in the analysis has two advantages. First, it allows to simultaneously analyze the average effect of the treatment and the interaction with the latent covariate. Second, the latent variable may reduce the variance of the structural residual, yielding more power to detect an average treatment effect.

It is expensive and time-consuming to carry out interventional studies and particularly so with a large number of participants. It is therefore important to know the minimum number of participants that can be used to answer the research questions. For the design of an interventional study, it is critical to estimate power to detect certain effects, such as average effects and interaction effects. The following section of this thesis aims to present some relevant power results for randomized interventional studies. The power estimation of the outlined general latent variable framework with regard to average treatment effects is based on a method developed by Satorra and Saris (1985).

The estimation of power to detect misspecified latent variable models has been discussed in Satorra and Saris (1985) and Saris and Satorra (1993); see also Saris and Stronkhorst (1984). In principle, power can be estimated for any model by carrying out a Monte Carlo study that records the proportion of replications in which the incorrect model is rejected. Satorra and Saris proposed a method that gives a tremendous simplification over such a brute force approach. A key technique is based on the likelihood-ratio chi-square test for maximum-likelihood estimation of mean and covariance structure models such as the one given in Equations 3.12 and 3.13. This technique is applied in the following to estimate power to detect intervention effects in the two-group latent variable model discussed above. The Satorra-Saris approach is particularly suitable for the intervention setting given that power estimates are desired for very specific model misspecifications concerning absence of treatment effects.

Under multivariate normality for $\boldsymbol{y}^{(j)}$, $(N-1)F_{min}$, where F_{min} is the optimal value in Equation 3.14, is distributed asymptotically as a chi-square variable when the model in Equations 3.12 and 3.13 is correct. Satorra and

Saris (1985) showed that when the model is incorrect but not highly misspecified, $(N-1)F_{min}$ is asymptotically distributed as a non-central chi-square variable with a certain non-centrality parameter, which can be approximated by a two-step procedure. This procedure involves two models, one more general that is assumed correctly specified and one more restrictive that is misspecified.

In the given intervention setting, we are interested in the power to detect intervention effects and the more restrictive model sets the corresponding parameter(s) to zero. As a first step, the more general two-group latent variable model is estimated including the treatment effect(s). In a second step, the estimated mean vectors and covariance matrices from Step 1 are used in place of the corresponding sample statistics and analyzed by the more restrictive model that sets the treatment effect parameter(s) to zero. The value of $(N-1)F_{min}$ in this second step represents an approximation to the non-centrality parameter. Once this parameter has been obtained, the power can be computed from tables for non-central chi-square distributions as a function of the degrees of freedom and the α-level of the test (see, e. g., Saris & Stronkhorst, 1984). Listing B.1 on page 148 gives a short R program that computes the power in this way. The degrees of freedom refer to the number of treatment effect parameters.

Saris and Satorra (1993) point to simulation studies that indicate that this procedure for estimating power can be sufficiently accurate for practical purposes at small sample sizes. To verify the accuracy for the presented two-group latent variable model, several simulation studies were carried out. The proportion of the replications for which the t value of the treatment effect (parameter) exceeded its 5% critical value was recorded. This t value refers to the incorrect hypothesis of zero treatment effect (parameter). The

studies focused on power values close to .8, varying the sample size and several other parameters as well as the method to estimate and test average treatment effects. The estimated power values of the Satorra-Saris method were compared to the simulation studies.

In the following sections, the Satorra-Saris method for estimating power will be used to compute power curves as a function of sample size for a variety of parameter combinations for the latent two-group model shown in Figure 3.3. Parameter values were chosen to represent various treatment effect sizes. These values generate the mean vectors on covariance matrices that are used in the second step of the power method. The power curves will be shown for different cases of the latent two-group model. Many different situations are in principle of interest: We may have an experimental study with individuals randomized into treatment and control groups, or the study may be non-experimental with pre-existing differences measured by latent covariates; there may be interactions between the treatment and the latent covariates, or not. This chapter considers experimental (randomized) designs with latent covariates.

The calculation of power curves calls for a consideration of effect size (Cohen, 1988). In a traditional two-group t-test setting, effect size is typically defined as the treatment and control group difference in outcome means, divided by a standard deviation based on the pooled outcome variance. A small effect size is typically taken to be .2, a medium effect size .5, an a large effect size .8 (Cohen, 1988).

In the latent variable model setting, the definition of effect size is not as straightforward. First, although Cohen-type definitions concern manifest variables, treatment effects in the discussed models can also be expressed in terms of latent variables. For example, in the Figure 3.3 model, the treatment

effect may be expressed in terms of the mean difference of the observed or the latent outcome, as well as the average effect of the treatment as given in Equation 3.22 (see also Equation 3.7).

Second, if reporting Cohen-like effect sizes for manifest variables, the standard deviation could be based on the control group rather than pooling over the treatment and control groups. The control group provides the normative value, whereas the treatment group variance in part reflects the treatment effect. In this thesis, I report effect sizes in several of these metrics.

3.5.1 Analysis of Examples

The power calculations to be illustrated below raise the issue of how small the sample size can be for trustworthy analysis results given the dependence on asymptotic theory. Here, it should be noted that considerations of power may suggest sample sizes that are smaller than what can be recommended for obtaining good estimates of parameters and standard errors. For example, the simple latent variable model in Figure 3.3 has 13 parameters in the control group (given tau-congeneric measurement models and the scaling described in section 3.4.1). A conventional requirement in the latent variable literature is 5 to 10 observations per parameter (see, e. g., Bentler & Chou, 1988). Using this rule of thumb would lead to a minimum of 65 preferably 130 control group observations. With a balanced design, a total of 130 to 260 control and treatment group observations may therefore be desired for this particular model. This total sample size requirement may exceed the number required for a power of at least .8 and this should be kept in mind when studying the power figures below.

For the power calculations of the latent two-group model given in Figure 3.3 (Equations 3.16 to 3.17), the following parameter values are chosen.

Data are generated using a normally distributed ξ with mean of zero

$$E(\xi) = 0 \tag{3.30}$$

and variance of one (for both groups)

$$Var(\xi) = 1. \tag{3.31}$$

Observations are randomly assigned to the two groups with fixed group sizes.

The parameters of the measurement models of the two latent variables η and ξ are set equal across the groups. $\lambda_{Z_k}^{(j)}$, the slopes (or factor loadings) for the measurement model of ξ, are set to one. The corresponding residual variances of the factor indicators are .5 with zero correlation between the two residuals ϵ_{Z_1} and ϵ_{Z_2}. The corresponding measurement intercepts ν_{Z_k} are zero. These parameters are chosen to give indicator reliabilities of about .66 using the following formula[4]

$$Rel(Z_k) = \frac{{\lambda_{Z_k}^{(j)}}^2 \; Var^{(j)}(\xi)}{{\lambda_{Z_k}^{(j)}}^2 \; Var^{(j)}(\xi) + \; Var^{(j)}(\epsilon_{Z_k})}, \tag{3.32}$$

for $k = 1, 2$ and $j = 1, 2$.

The parameters of the measurement model of η are set to the same values as the measurement model of ξ yielding similar reliabilities of Y_k. However, the reliability of the factor indicators for η varies to some extend with the variance of η, which depends on the parameters of the structural equations. The parameters of the measurement models are the same throughout the following power calculations and are summarized in the following equations

$$\nu_{Y_k}^{(j)} = \nu_{Z_k}^{(j)} = 0, \quad \lambda_{Y_k}^{(j)} = \lambda_{Z_k}^{(j)} = 1, \text{ and} \tag{3.33}$$

$$Var^{(j)}(\epsilon_{Y_k}) = \; Var^{(j)}(\epsilon_{Z_k}) = 0.5, \text{ where } k = 1, 2. \tag{3.34}$$

[4]This formula is derived from Bollen (2002) when there are no correlated errors of measurement.

The measurement residuals are uncorrelated with each other and uncorrelated with the latent variables.

The structural intercept given the control group is set to

$$\alpha^{(1)} = 0.2. \tag{3.35}$$

Given $E(\xi) = 0$ and $\alpha^{(1)} = 0.2$, the average effect is computed by the difference between the two structural intercepts (see Equation 3.22)

$$AE = \alpha^{(2)} - 0.2. \tag{3.36}$$

The structural intercept in group two varies between the values .2, .4, .7, and 1 in order to vary the average treatment effect. The corresponding average treatment effect values are 0, .2, .5, and .8.

The group-specific variances of the structural residual, are held equal across the groups: $Var^{(j)}(\zeta) = Var(\zeta)$. The structural residual variance is varied in order to vary the proportion of the variance of η that is explained by ξ. $Var(\zeta)$ varies between the values .1, .3, .5, .7, .9, and 1. The structural slope in the first group is set depending on $Var(\zeta)$ by the formula

$$\beta^{(1)} = \sqrt{\frac{1 - Var(\zeta)}{Var(\xi)}}. \tag{3.37}$$

With $Var(\xi) = 1$, $Var(\eta)$ the variance of the latent outcome in the first group is always one (see Equation C.4 on page 151). As a consequence the correlation between ξ and η in the first group is equal to $\beta^{(1)}$ (see Equation C.5 on page 151). The corresponding coefficient of determination for group one is

$$R^{2\ (1)}_{\eta\,|\,\xi} = \frac{Var(E_{X=1}(\eta\,|\,\xi))}{Var^{(1)}(\eta)} = {\beta^{(1)}}^2. \tag{3.38}$$

Table 3.1 on the following page gives a summary about the parameters that vary with $Var(\zeta)$. The structural slope of the control group is computed

Table 3.1

The Values for $Var(\zeta)$ and the Corresponding Values of $\beta^{(1)}$ (see Equation 3.37) and $R^{2\ (1)}_{\eta\,|\,\xi}$.

$Var(\zeta)$	1	.9	.7	.5	.3	.1
$\beta^{(1)}$	0	$\sqrt{.1}$	$\sqrt{.3}$	$\sqrt{.5}$	$\sqrt{.7}$	$\sqrt{.9}$
$R^{2\ (1)}_{\eta\,|\,\xi}$	0	.1	.3	.5	.7	.9

so that the variance of η in the control group is always one. This is done because the variance of η in the control group serves as a norm used to compute the measures of effect size for the average treatment effect as well as the interaction effect.

Following Cohen's (1988) classification, the average treatment effect values of 0, .2, .5 and .8 are referred to as a small, medium and large effect, respectively. The interaction effects are varied by adding 0, .2, .5, or .8 to $\beta^{(1)}$. The non-zero interaction effects are categorized as small, medium and large correspondingly. The last examples of the power analyzes for randomized designs will focus on unbalanced designs and vary the proportion of observations in the treatment group.

3.5.2 Interaction Effects

An aspect of the latent multi-group model is to estimate and test interaction effects between treatment and latent covariates. The following section describes how to analyze the power of the latent multi-group model to detect interaction effects with different effect sizes. Power values are obtained using the Satorra-Saris method described in section 3.5. As a first step, the mean vectors and covariance matrices for each group are computed using

Equations 3.12 to 3.13 on page 58 for the latent two group model with the population parameters as specified above. The resulting mean vectors and covariance matrices are written in a file (`pop.dat`) so that Mplus can read them.

In a second step, the more general two-group latent variable model (including the interaction) as well a the restricted model (without the interaction) are "estimated". The Mplus input for the more general model is given in Listing A.6 on page 144. The Mplus input of the model restricted for no interaction effect is obtained from the same input by commenting out line 11 and uncommenting line 12. The number of observations for each group is set arbitrary to 1000, a value large enough for sufficient precision.

The third step is to compute the difference between the chi-square value of the restricted model and the chi-square value of the unrestricted model. This chi-square difference is then used in the R program to compute power values (see Listing B.1). The power values for other sample sizes are calculated by multiplying the chi-square difference by the ratio of the new sample size to the original sample size (here 2000). This value represents an approximation to the non-centrality parameter. The restriction for the interaction effect has $df = 1$. An alpha level of 0.05 is used throughout this discussion.

Figure 3.4 gives power curves for the two-group latent variable model to detect interaction effects with different effect sizes. The sample sizes range from 50 to 1,000. Here, sample size refers to the total number of observations that are assigned equally to control and treatment resembling a balanced design.

The variance of the structural residual is fixed to $Var(\zeta) = 1$ and the corresponding value of $\beta^{(1)}$ to 0. In other words, the latent covariate does not explain any variance in the control group (i. e. $R^{2\ (1)}_{\eta\,|\,\xi} = 0$; see Table 3.1).

Note again that the variance of η in the control group, $Var^{(1)}(\xi) = 1$ is used as a norm for the effect size measures of the interaction effect. Three values for the interaction effect are considered .2, .5, and .8 with the corresponding effect size small, medium and large.

The structural intercept is set to $\alpha^{(j)} = 0.2$ in both groups, resulting in an average treatment effect of zero. It is important to note that the power analysis of the interaction effect is not limited to the case of no average treatment effect. This means that it is possible to assess the power of detecting an interaction while allowing for an average treatment effect.

Figure 3.4 shows that a large total sample size of over 1,000 is needed to achieve a power of .8 for a small interaction effect size. The sample sizes needed to achieve the same power is considerably smaller for medium or large interaction effects (about 200 and less than 100 respectively). This finding is in line with the case of interactions in multiple regression (Aiken & West, 1991) and longitudinal modeling (B. O. Muthén & Curran, 1997).

A second set of power curves (with regard to the interaction effect) is computed for the case that ξ explains 90% of the variance of η in the control group. The structural residual variance is set to $Var(\zeta) = 0.1$ and the corresponding structural slope in group one to $\beta^{(1)} = \sqrt{0.1}$. This yields a coefficient of determination of $R^{2\ (1)}_{\eta\,|\,\xi} = 0.9$ (see Table 3.1). All other parameters are as in the previous example. Figure 3.5 shows power curves of the latent two group model to detect an interaction with a small, medium, and large effect size.

Given a much smaller variance of the structural residual it is obvious that less observations are needed to achieve the same power as before in Figure 3.4. A power of .8 is achieved with about 780 observations given a small interaction, with about 150 observations given a medium interaction

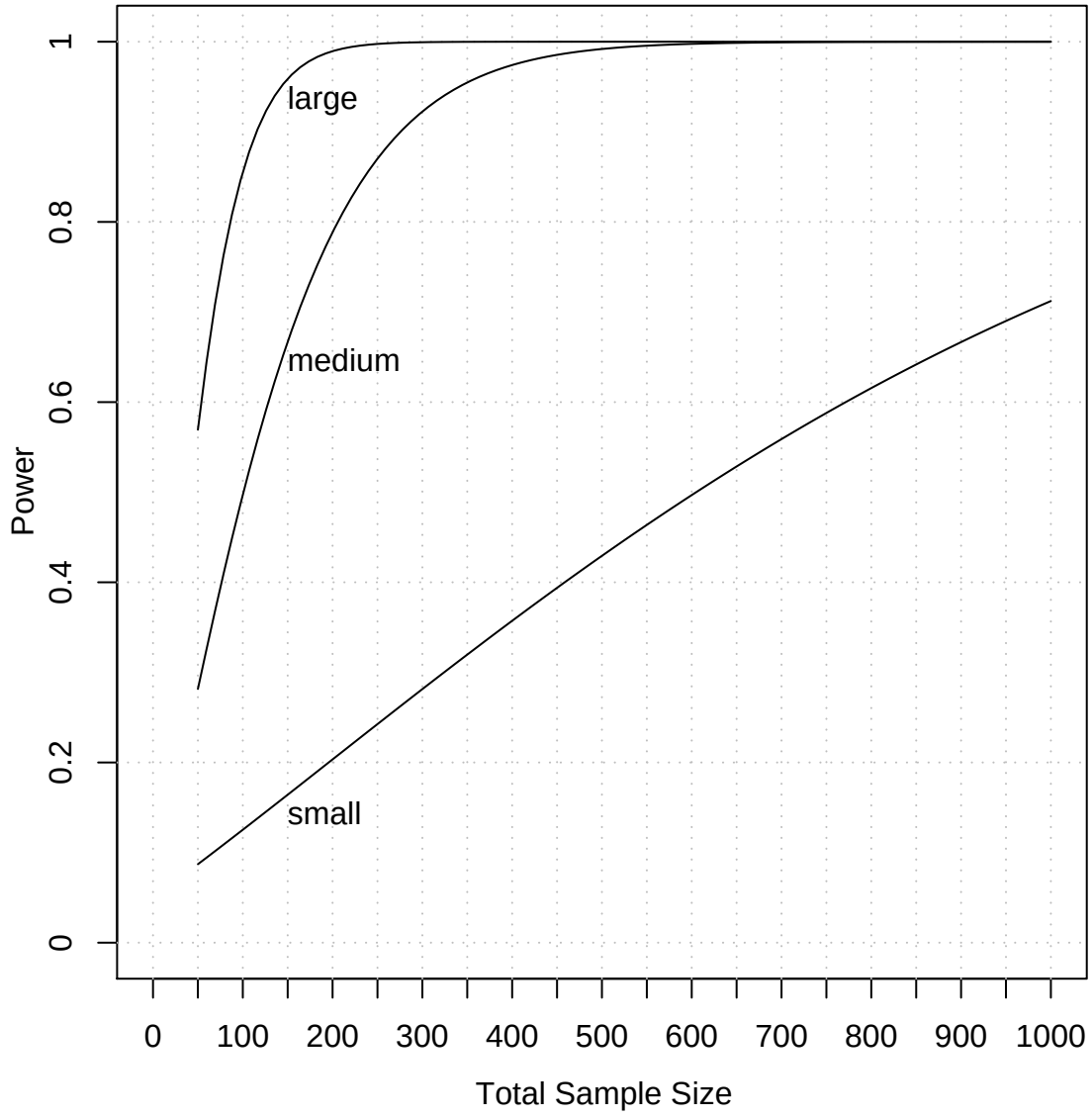


Figure 3.4: Power to detect a small, medium, or large interaction effect as a function of total sample size for a balanced randomized design, no average treatment effect, and $R^{2\ (1)}_{\eta\,|\,\xi} = 0$.

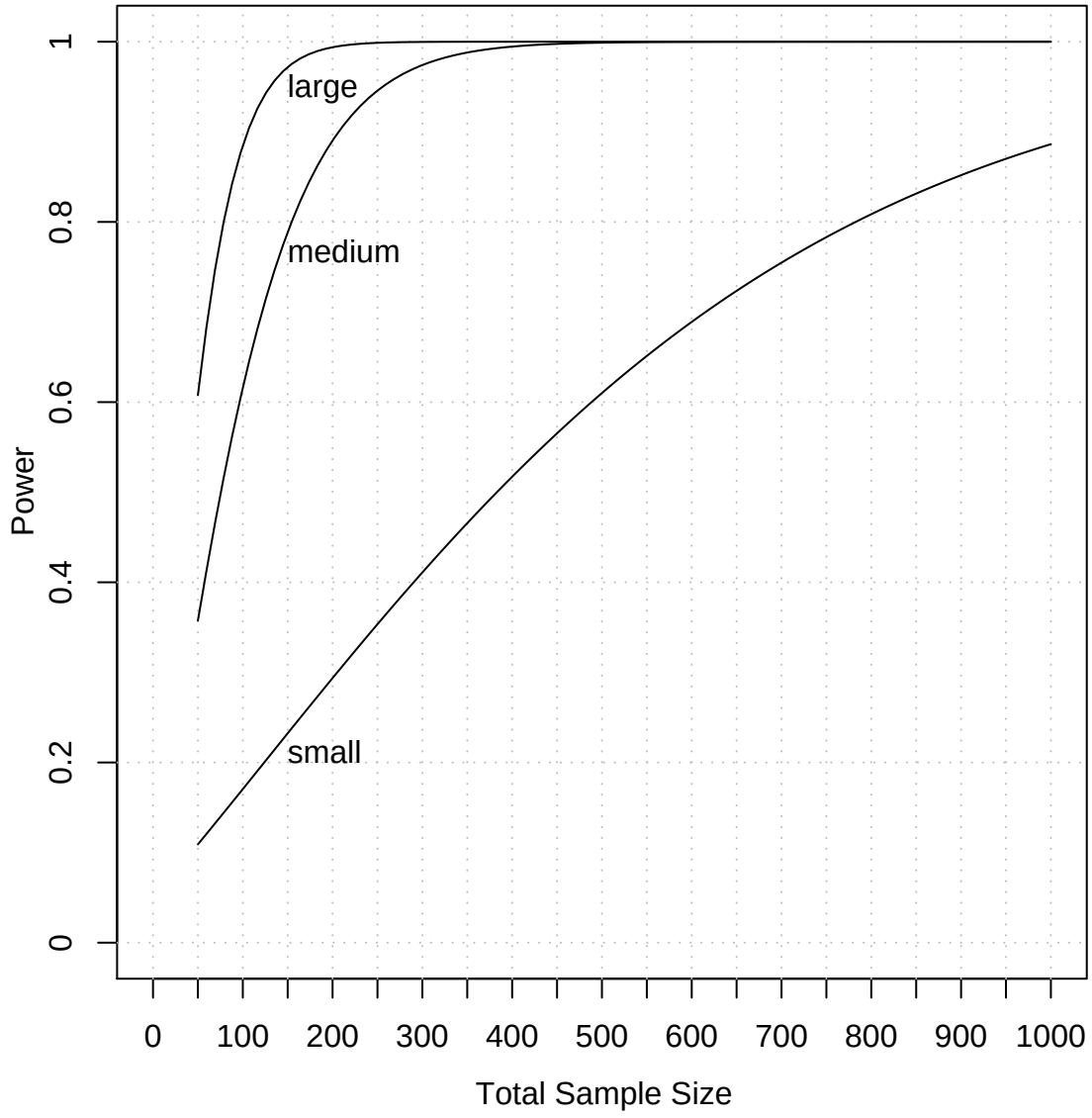


Figure 3.5: Power to detect a small, medium, or large interaction effect as a function of total sample size for a balanced randomized design, no average treatment effect, and $R^{2\ (1)}_{\eta\,|\,\xi} = 0.9$.

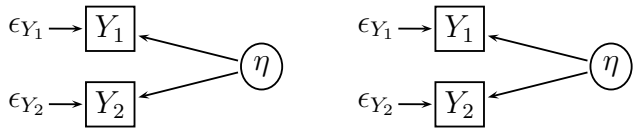
and with less than 100 observations given a large interaction effect. It is still remarkable how many observations are required to detect small interactions, even if the structural residual variance is small.

To verify the Satorra-Saris method for the given model a Monte Carlo Study with a balanced and randomized design is conducted with a sample size of 800, $R^{2\ (1)}_{\eta\,|\,\xi} = .9$, no average treatment effect and a small interaction effect. The mean of $1,000$ estimates for the interaction was .2 which is exactly the expected value, the standard deviation of these estimates was .08. The observed power of .709 was obtained by counting the number of significant interactions (709 out of 1000) at an alpha level of .05. This number is close to the value .703 computed with the Satorra-Saris method. The Mplus input for this Monte Carlo simulation is given in Listing A.8 on page 145.

3.5.3 Average Treatment Effects

The main focus of the power analysis is now on the average treatment effect. The following paragraphs discuss the power of the latent multi-group approach to detect an average treatment effect for randomized studies. As mentioned earlier, an unbiased estimate for the average treatment effect may be obtained simply by the mean difference in the (latent or observed) outcome variables. Figure 3.6 shows the pathdiagram of a latent multi-group model without the latent covariate.

This latent multi-group model can be identified in a similar way as the latent multi-group model that includes the latent covariate, using tau-congeneric measurement models and setting the mean of the latent outcome in one group to zero. However, constraining this model so that the average treatment effect is set to zero $\alpha^{(2)} - \alpha^{(1)} = 0$, yields a non-identified model. Given the



(a) Group 1 (b) Group 2

Figure 3.6: A multi-group model without the latent covariate.

tau-congeneric measurement model, the parameter $\lambda_{Y_2}^{(1)}$ is identified with

$$\lambda_{Y_2}^{(1)} = \frac{E(Y_2)^{(1)} - E(Y_2)^{(2)}}{\alpha^{(1)} - \alpha^{(2)}}. \tag{3.39}$$

Given the restriction for no average treatment effect $\alpha^{(1)} - \alpha^{(2)} = 0$, it is obvious that $\lambda_{Y_2}^{(1)}$ is not identified. For a power analysis of sample data with the sample means and covariances deviating from the population values the restricted model might be estimable. For power analysis in this dissertation, that are based on population parameters, essentially tau-equivalent measurement models are used to avoid identification problems that occur with the average treatment effect constrained to zero. This identification issue of the model without the latent covariate will be discussed again in the Monte Carlo studies below.

The model ignoring the latent covariate is compared to the model including the latent covariate with regard to the power to detect an average treatment effect. This model comparison is done for an average treatment effect with different effect sizes and with the latent covariate explaining different proportions of the latent outcome variance. As described above the variance of the structural residual is varied in order to vary the proportion of variance of the latent outcome that is explained by the latent covariate in group one. The measure used to report this dependency of η on ξ is again the coefficient of determination $R^{2\ (1)}_{\eta\,|\,\xi}$ (see Table 3.1).

Several parameter settings will be discussed. They include the case without and with an interaction effect present as well as settings for unbalanced designs, where the number of observations differ between the groups. The power of the latent multi-group model to detect an average treatment effect will be analyzed for the case that the multi-group model incorporates the interaction effect as well as the case where interactions are ignored even though it is present. For certain parameter constellations it will also be necessary to consider the essentially tau-equivalent measurement model because the more general tau-congeneric model is not identified (as mentioned above). Tau-equivalent measurement models are more restricted because the measurement slopes for a latent variable are set equal. This issue is discussed whenever appropriate.

The general Mplus input for the power analysis of the model including the latent covariate used in this section is given in Listing A.6 on page 144. Several sub-versions of this model are obtained by commenting or uncommenting lines. As mentioned above the interaction is set to zero by commenting out line 11 and uncommenting line 12 (i. e. restricting the structural slopes to be equal across groups). The average treatment effect is set to zero by uncommenting line 13. Commenting out line 8 and uncommenting line 9 changes the measurement models from tau-congeneric to essentially tau-equivalent for both latent variables (in all groups).

The Mplus input for the latent multi-group model ignoring the latent covariate is given in Listing A.7 on page 145. The corresponding restricted model is obtained by uncommenting line 11. Commenting out line 10 and uncommenting line 9 changes the measurement models from essentially tau-equivalent to tau-congeneric for both latent variables (in all groups).

3.5.4 No Treatment-Covariate Interactions

This section compares the latent multi-group model including the latent covariate to the model ignoring the latent covariate with regard to the power to detect an average interaction effect. The cases considered here include the balanced randomized design with no interactions between treatment and latent covariate. The effect size of the average treatment effect is varied as well as the latent outcome variance explained by the latent covariate.

Consider the case where the latent covariate does not explain any variance of the latent outcome (i. e., $\beta^{(j)} = 0$). Given this case the latent multi-group model involving the latent covariate is not identified for tau-congeneric measurement models, hence, essentially tau-equivalent measurement models are used.

Based on the Satorra-Saris method, chi-square difference values are computed for the model without the latent covariate, with the latent covariate as well as for the model with the latent covariate but the interactions constrained to zero. The resulting chi-square difference values are identical for all three models: 15.95 for a small average effect, 97.68 for a medium average effect, and 241.12 for a large average effect. Hence, all models yield the same power for the given parameters.

To verify these results, a Monte Carlo Study is conducted with $1,000$ replications each with a sample size of $1,000$ (500 observations in each group), with an average effect set to .2. The data is generated with the Mplus input given in Listing A.9 on page 146. This input also includes the instructions for the model including the latent covariate, the interaction and tau-congeneric measurements. The Mplus input for the model without the latent covariate (using the same data) is given in Listing A.10 on page 147. Several other models are also estimated, varying the measurement models and whether the

Table 3.2

Monte Carlo Outcomes for the Power to Detect a Small Average Effect Given no Interaction, $R^{2\ (1)}_{\eta\,|\,\xi} = 0$*, and a Sample Size of* 1000.

Model	M	SD	rf_e	rf_o	C	W
1. w/ ξ, TC, w/ int.	0.192	0.073		0.737	699	1076
2. w/ ξ, TC, w/o int.	0.198	0.074		0.734	721	1045
3. w/ ξ, TE, w/ int.	0.199	0.071	0.806	0.817	1000	0
4. w/ ξ, TE, w/o int.	0.199	0.071	0.806	0.817	1000	0
5. w/o ξ, TC	0.199	0.077		0.725	991	468
6. w/o ξ, TE	0.199	0.071	0.806	0.819	1000	0
7. Mean Diff.	0.199	0.069		0.816	1000	0

Note. Models: w/ or w/o ξ = with or without the latent covariate; TC or TE = tau-congeneric or tau-eqivalent measurement models for all latent variables; w/ or w/o int. = with or without the interaction; Mean Diff. = two sample t-test of $(Y_1 + Y_2)/2$; M, SD = means and standard deviation of the $1,000$ average effect estimates; rf_e = expected relative frequency of significant tests based on the Satorra-Saris method; rf_o = observed relative frequency of significant tests; C = number of completed models; W = number of warnings.

interaction is set to zero or not.

As mentioned above it is possible to get an causally unbiased estimate of the average effect simply by the mean difference of the observed outcome variables. This model is included in the study and named "Mean Diff.". The average treatment effect is analyzed with a two sample t-test of $(Y_1 + Y_2)/2$. The results for all models are given in Table 3.2.

Recorded are the mean and the standard deviation of the $1,000$ average treatment estimates of each model. The means are very close to the true

parameter .2, indicating unbiasedness for all models. The standard deviations are similar which indicates similar accuracy of all models.

The proportion of the replications for which the t-value of the average treatment effect exceeds its 5% critical value is referred to as rf_o. This t-value refers to the incorrect hypothesis of zero average treatment effect. The study focused on power values close to .8. Due to identification issues (as discussed above) the Satorra-Saris method is applicable only for the latent variable models with essentially tau-equivalent measurement models. A good agreement is obtained for all three models. The observed relative frequencies (rf_o) are close to the power of .86 obtained by the Satorra-Saris method (rf_e).

Although the latent variable models with tau-congeneric measurements are not identified for the (true) population parameters, they might be identified if the sample parameters deviate from the population parameters. However, estimation is not completed for all samples, and even if the model estimation is completed a large number of warnings indicate caution when interpreting the results for the latent variable models with tau-congenereric measurements.

The model estimating the average treatment effect solely with the observed outcome variables ("Mean Diff.") has no advantage with regard to the power. However, it is based on less restrictive assumptions than the latent variable models. To summarize the results for the case that the latent covariate has no effect on the latent outcome, all the models (whether they include the latent covariate or not, as well as the model based on the observed outcome variable) detect an average treatment effect with the same power. In other words, if the assumptions of the latent variable model are met, it does not hurt to include a latent covariate in the model, even if this

latent covariate does not explain any variance in the latent outcome.

It is of interest, what gain in power to expect if the latent covariate does actually explain variance of the latent outcome. The following power analysis therefore considers the case that the latent covariate explains a certain proportion of the latent outcome variance. The measure that is used to reflect this proportion is again $R^{2\ (1)}_{\eta\,|\,\xi}$, the coefficient of determination in group one.

As mentioned above the parameters that are varied in order to vary $R^{2\ (1)}_{\eta\,|\,\xi}$ are $Var^{(j)}(\zeta)$ and $\beta^{(1)}$ (see Table 3.1). For a balanced design with no interaction present the Satorra-Saris method is used to analyze the power of the latent variable model without ξ in comparison to the latent variable model including ξ.

The first result is that the power of the latent model without ξ to detect a given average effect is the same for all values of $R^{2\ (1)}_{\eta\,|\,\xi}$. With $\beta^{(1)} \neq 0$ the latent variable model including ξ is identified with tau-congeneric measurement models. The second result is that all[5] latent variable models including ξ share the same power to detect an average treatment effect for a given value of $R^{2\ (1)}_{\eta\,|\,\xi}$.

The third and most important finding is that the power of latent models including ξ increases significantly the more outcome variance ξ explains (i. e. the larger $R^{2\ (1)}_{\eta\,|\,\xi}$). The power curves in Figure 3.7 show the power of the latent two-group model to detect a small average treatment effect ($\alpha^{(2)} - \alpha^{(1)} = .2$) as a function of sample size as well as $R^{2\ (1)}_{\eta\,|\,\xi}$. Here, sample size refers to the total number of individuals assigned equally to the control and treatment group (balanced case).

Given the case that the latent covariate does not explain any latent outcome variance (i. e. $\beta^{(1)} = 0$, $R^{2\ (1)}_{\eta\,|\,\xi} = 0$), the model including ξ and the

[5]Models estimating the interaction or setting it to zero as well as models with essentially tau-equivalent or tau-congeneric measurement models.

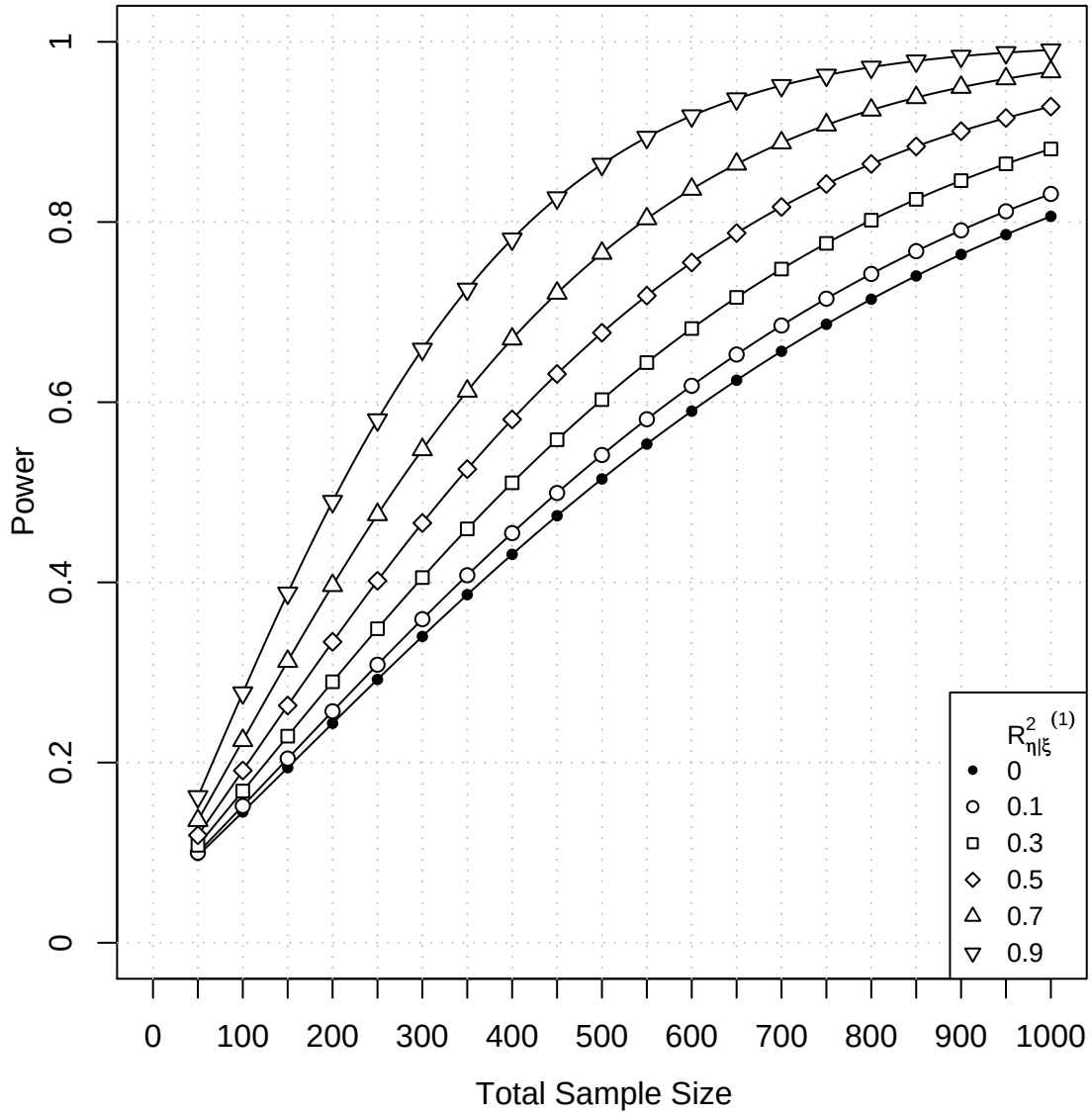


Figure 3.7: The power of the latent two-group model to detect a small average treatment effect as a function of total sample size and $R^{2\ (1)}_{\eta\,|\,\xi}$ for a balanced design and no interaction between treatment and latent covariate.

model ignoring ξ yield exactly the same power, represented by the bottom curve in Figure 3.7. This curve also represents the power of the latent variable model ignoring ξ for any other value of $R^{2\ (1)}_{\eta\,|\,\xi}$. The bottom curve shows that a large total sample size of about 1000 is needed to achieve a power of .8 if ξ is ignored or it is included but does not explain any latent outcome variance.

The five curves above this bottom curve represent the power of the latent variable model including ξ for cases where ξ does explain latent outcome variance. Figure 3.7 reveals a gain in power for the model including ξ in comparison to the model without ξ: the larger the proportion of explained outcome variance the larger this gain in power. Considering the case where ξ explains 90% of the latent outcome variance, the total sample size required to achieve a power of .8 for the model including ξ is less than have the sample size required to achieve the same power with a model ignoring ξ. Similar results may be obtained for different effect sizes of the average treatment effect. Figure 3.8 shows the power curves of a medium average effect ($\alpha^{(2)} - \alpha^{(1)} = 0.5$).

Monte Carlo studies are conducted to verify the results for the balanced case with no interaction present. A total sample size of 400 (200 in each group) is chosen. The average treatment effect is set to .2 (i. e. a small effect size) and $R^{2\ (1)}_{\eta\,|\,\xi}$ varied between the values .1, .5, and .9. The same models are used as in the previous Monte Carlo study.

The outcomes are given in Table 3.3. Comparing the means of the estimates for the average treatment effect to its true population value (.2) reveals that all models provide an unbiased estimate. The standard deviation of all models ignoring ξ is nearly the same. Comparing the standard deviation between the models and the three values of $R^{2\ (1)}_{\eta\,|\,\xi}$ reveals that the models

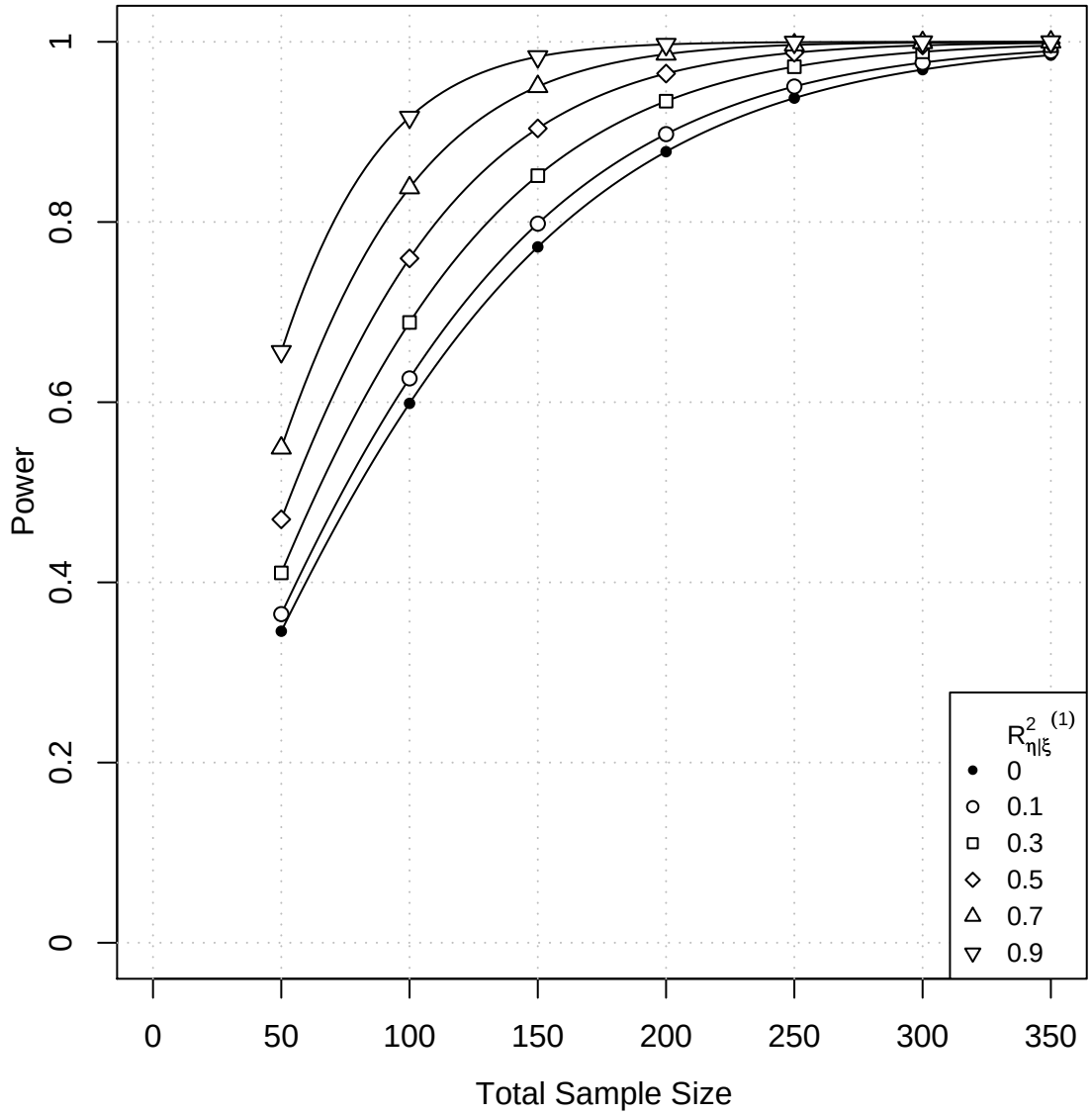


Figure 3.8: The power of the latent two-group model to detect a medium average treatment effect as a function of total sample size for a balanced design and no interaction between treatment and latent covariate.

Table 3.3

Results for the Monte Carlo Study for the Power to Detect a Small Average Treatment Effect for the Case That $R^{2\ (1)}_{\eta\,|\,\xi} = 0.1$, 0.5, *or* 0.9 *and a Total Sample Size of* 400.

Model	M	SD	rf_e	rf_o	C	W
			$R^{2\ (1)}_{\eta\,|\,\xi} = 0.1$			
1. w/ ξ, TC, w/ int.	0.199	0.107	0.455	0.463	1000	197
2. w/ ξ, TC, w/o int.	0.199	0.107	0.455	0.466	1000	176
3. w/ ξ, TE, w/ int.	0.199	0.107	0.455	0.460	1000	0
4. w/ ξ, TE, w/o int.	0.199	0.107	0.455	0.461	1000	0
5. w/o ξ, TC	0.212	0.122		0.424	938	796
6. w/o ξ, TE	0.200	0.111	0.431	0.450	1000	0
7. Mean Diff.	0.200	0.109		0.448	1000	0
			$R^{2\ (1)}_{\eta\,|\,\xi} = 0.5$			
1. w/ ξ, TC, w/ int.	0.197	0.092	0.581	0.574	1000	0
2. w/ ξ, TC, w/o int.	0.197	0.092	0.581	0.573	1000	0
3. w/ ξ, TE, w/ int.	0.197	0.092	0.581	0.573	1000	0
4. w/ ξ, TE, w/o int.	0.197	0.092	0.581	0.573	1000	0
5. w/o ξ, TC	0.204	0.122		0.364	944	776
6. w/o ξ, TE	0.196	0.111	0.431	0.420	1000	0
7. Mean Diff.	0.196	0.112		0.410	1000	0
			$R^{2\ (1)}_{\eta\,|\,\xi} = 0.9$			
1. w/ ξ, TC, w/ int.	0.203	0.072	0.781	0.806	1000	167
2. w/ ξ, TC, w/o int.	0.203	0.072	0.781	0.806	1000	138
3. w/ ξ, TE, w/ int.	0.203	0.072	0.781	0.806	1000	171
4. w/ ξ, TE, w/o int.	0.203	0.072	0.781	0.805	1000	150
5. w/o ξ, TC	0.218	0.122		0.406	956	728
6. w/o ξ, TE	0.207	0.111	0.431	0.453	1000	0
7. Mean Diff.	0.207	0.110		0.441	1000	0

Note. See Table 3.2 for a description of the abbreviations.

including the latent covariate estimated the average treatment effect more precisely. The more latent outcome variance is explained by ξ the larger this increase in precision.

Given $R^{2\ (1)}_{\eta\,|\,\xi} = .5$ estimation problems and warning occurred only for the latent variable model ignoring ξ with tau-congeneric measurements. As discussed earlier this is the only model that is not identified for zero average treatment tests. It can also be seen that for values of $R^{2\ (1)}_{\eta\,|\,\xi} = .1,$ and .9 the latent variable models with tau-congeneric measurements produce warnings. And for $R^{2\ (1)}_{\eta\,|\,\xi} = .9$ even the essentially tau-equivalent measurement models produce some warnings.

In summary the Monte Carlo studies show sufficient agreement between the expected and observed proportion of significant tests for all models indicating that the Satorra-Saris method is applicable for the considered examples in which ξ explains latent outcome variance.

3.5.5 Interactions

The main focus of this dissertation is on models that include interactions. In this section the Satorra-Saris method is applied in order to show how to analyze the power of the latent multi-group model to detect an average treatment effect in randomized designs if an interaction effect between treatment and latent covariate is present. The focus will be on interaction effects with small effect size because this effect size is regarded to be the most relevant in applied research.

Even with an interaction effect between latent covariate and the treatment present, it is possible to estimate and test the average treatment effect with models that ignore the latent covariate (see Figure 3.6). This model is again compared to models that include the latent covariate. It is important to keep

in mind that the focus of this section is on the power to detect an average treatment effect and not to detect an interaction effect. If the focus of a study is on the interaction effect as well as the average treatment effect, the power of both effects will have to be considered when choosing the total sample size.

The model comparison with regard to the power of detecting a small average treatment effect is again performed with the Satorra-Saris method by choosing the same parameters as in the last section. The difference here is that the interaction effect is non-zero and set to $\beta^{(2)} - \beta^{(1)} = .2$. The average treatment effect is set to $\alpha^{(2)} - \alpha^{(1)} = .2$ and $R^{2\ (1)}_{\eta\,|\,\xi}$, the proportion outcome variance explained by ξ, is again varied according to Table 3.1.

The results of the Satorra-Sarris power analysis are shown in Figure 3.9. Power curves are given for the latent two-group model involving the interaction between treatment and ξ (and tau-congeneric measurements). The different curves refer to the proportion of outcome variance explained by ξ in the control group. The more outcome variance ξ explains, the larger the power in order to detect a small average treatment effect.

The additional curve marked with the solid triangle represents the power of the latent variable model ignoring the latent covariate given $R^{2\ (1)}_{\eta\,|\,\xi} = .9$. The power curves for this model given the other values of $R^{2\ (1)}_{\eta\,|\,\xi}$ are omitted. All of these omitted curves are slightly above the curve given but below the lowest curve of the model that includes ξ. The results of the power analysis for the given parameters reveals similar results as in the last section where the interaction effect was set to zero. Given a non-zero interaction the model that includes the latent covariate detects an average treatment effect with more power than a model that ignores ξ. The more of the variance of the latent outcome ξ explains the larger is the gain in power that results from

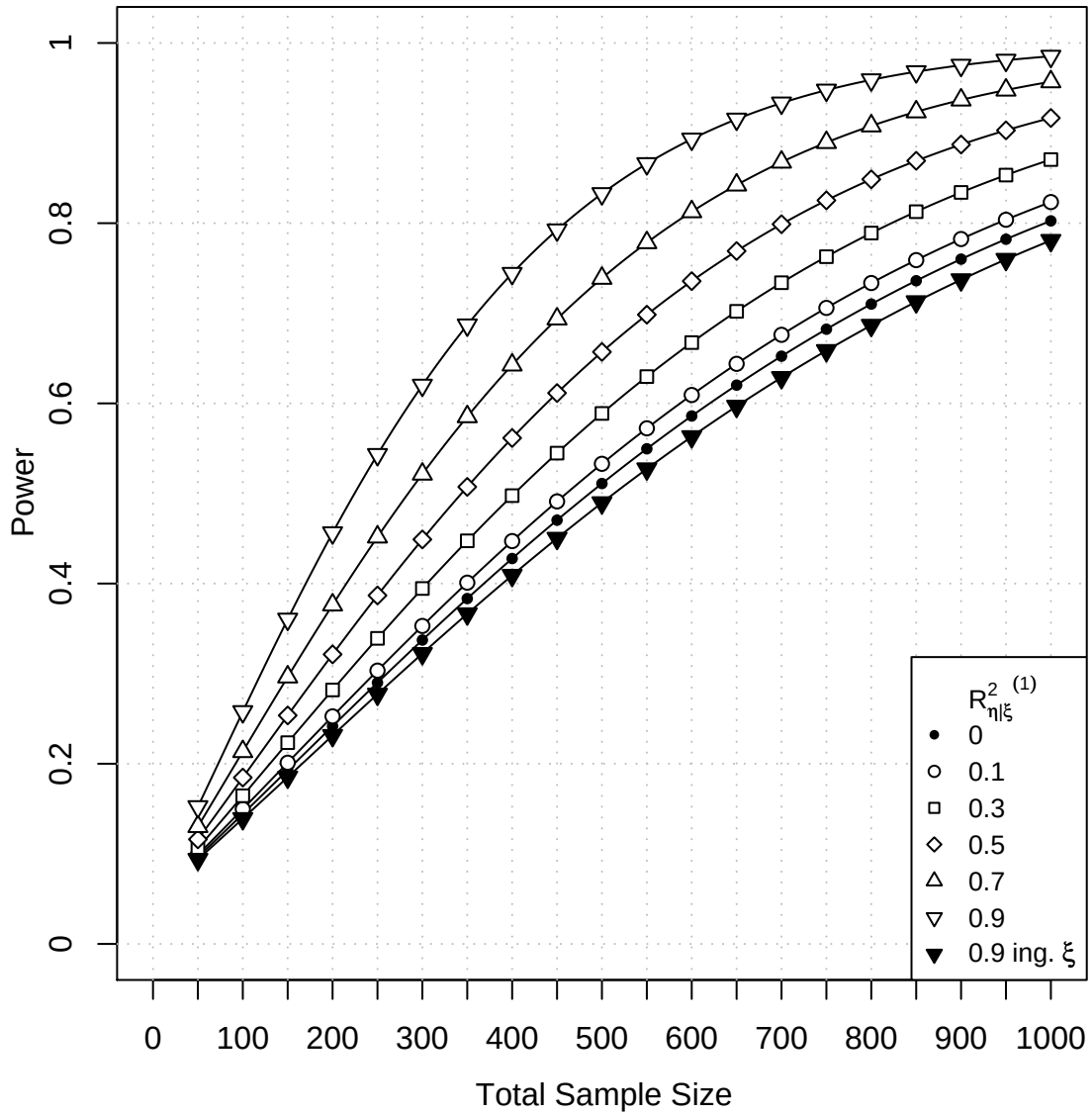


Figure 3.9: The power of the latent two-group model to detect a small average treatment effect as a function of total sample size and $R^{2\ (1)}_{\eta\,|\,\xi}$ for a balanced design and a small interaction between treatment and latent covariate. The curve marked with the solid triangle represents the power of the latent variable model ignoring the latent covariate given $R^{2\ (1)}_{\eta\,|\,\xi} = .9$.

including ξ in the model. The total sample size required to achieve a power of .8 is again less than half for the model including ξ compared to the model without ξ given $R^{2\ (1)}_{\eta\,|\,\xi} = .9$.

Monte Carlo studies are conducted to verify the results of the Satorra-Saris power analysis. The parameters are set to the same values as just described for Figure 3.9. The average treatment effect is set to $\alpha^{(2)} - \alpha^{(1)} = .2$, the interaction is set to $\beta^{(2)} - \beta^{(1)} = .2$ and $R^{2\ (1)}_{\eta\,|\,\xi} = 0$, .5, or .9. A total sample size of 600 is chosen with observations again distributed equally to the two groups. Estimation for the same models as in the previous Monte Carlo Studies are recorded. For a given parameter set the estimations are based on the same data for each model.

The outcomes are given in Table 3.4. Comparing the means of the estimates for the average treatment effect to the true population value .2 reveals that all models provide an unbiased estimate. Comparing the standard deviation between the models that include ξ to all the models that ignore ξ for the three values of $R^{2\ (1)}_{\eta\,|\,\xi}$ reveals that the models including ξ estimate the average treatment effect more precisely. The more of the variance of the latent outcome ξ explains the larger is the increase in precision.

Comparing the expected and observed proportion of significant tests for all models indicates that the Satorra-Saris method is applicable all considered examples with interaction effects present. The latent variable model ignoring ξ with tau-congeneric measurement model is not identified for the population parameters. Consequently, no value for the expected proportion of significant tests is given. Based on the $1,000$ samples this model again produces a large number of warnings and some uncompleted model estimations. The results for this model have to be interpreted with caution. Given $R^{2\ (1)}_{\eta\,|\,\xi} = 0$ the models with tau-congeneric measurement also show a large number

Table 3.4

Results for the Monte Carlo Study for the Power to Detect a Small Average Treatment Effect for the Case that $R^{2\ (1)}_{\eta\,|\,\xi} = 0$, .5, *or* .9, *a Small Interaction and a Total Sample Size of* 600.

Model	M	SD	rf_e	rf_o	C	W
			$R^{2\ (1)}_{\eta\,\|\,\xi} = 0$			
1. w/ ξ, TC, w/ int.	0.192	0.091	0.586	0.532	966	798
2. w/ ξ, TC, w/o int.	0.196	0.092	0.586	0.532	918	969
3. w/ ξ, TE, w/ int.	0.199	0.091	0.586	0.580	1000	0
4. w/ ξ, TE, w/o int.	0.199	0.091	0.586	0.581	1000	0
5. w/o ξ, TC	0.202	0.101		0.505	973	656
6. w/o ξ, TE	0.199	0.092	0.583	0.576	1000	0
7. Mean Diff.	0.199	0.091		0.568	1000	0
			$R^{2\ (1)}_{\eta\,\|\,\xi} = 0.5$			
1. w/ ξ, TC, w/ int.	0.202	0.077	0.736	0.739	1000	0
2. w/ ξ, TC, w/o int.	0.202	0.077	0.737	0.738	1000	0
3. w/ ξ, TE, w/ int.	0.202	0.077	0.736	0.741	1000	0
4. w/ ξ, TE, w/o int.	0.202	0.077	0.737	0.741	1000	0
5. w/o ξ, TC	0.205	0.105		0.496	966	756
6. w/o ξ, TE	0.199	0.097	0.539	0.521	1000	0
7. Mean Diff.	0.199	0.098		0.517	1000	0
			$R^{2\ (1)}_{\eta\,\|\,\xi} = 0.9$			
1. w/ ξ, TC, w/ int.	0.199	0.062	0.893	0.887	1000	88
2. w/ ξ, TC, w/o int.	0.199	0.062	0.895	0.889	1000	139
3. w/ ξ, TE, w/ int.	0.199	0.062	0.893	0.888	1000	97
4. w/ ξ, TE, w/o int.	0.199	0.062	0.895	0.888	1000	151
5. w/o ξ, TC	0.206	0.106		0.501	966	857
6. w/o ξ, TE	0.201	0.098	0.526	0.536	1000	0
7. Mean Diff.	0.201	0.100		0.533	1000	0

Note. See Table 3.2 for a description of the abbreviations.

of warnings and several uncompleted model estimations. For $R^{2\ (1)}_{\eta\,|\,\xi} = .9$ warnings are recorded for all latent variable models but all model estimations are completed.

To summarize the results of the power analysis for balanced randomized designs, it can be stated that the Satorra-Saris method provides sufficient agreement between the predicted power values and the results of the Monte Carlo studies. The examples included the case without and with an interaction effect present and models that include and ignored the interaction.

3.5.6 Unbalanced Designs

This section discusses effects on power of deviations from balanced data. First, consider again the simple case without an interaction effect present and the latent covariate explaining no latent outcome variance. For this case, Figure 3.10 shows, how the power varies as a function of the proportion of treatment-group observations for a given total sample size of 250, 500, 750, and $1,000$. The power is the same for all latent models, whether they include or ignore ξ and whether they fix the interaction to zero or not. The power curves are symmetric around the balanced case where the proportion is .5. Choosing an unbalanced design in favor of more treatment or more control observations reduces the power to detect an average treatment effect.

Figure 3.11 shows the power of the latent variable model including ξ to detect a small average treatment effect for the case that the latent covariate explains 90% of the latent outcome variance in the control group ($R^{2\ (1)}_{\eta\,|\,\xi} = 0.9$) and no interaction present. The power of the latent variable model ignoring ξ for this case is the same as shown in Figure 3.10. Comparing the model including ξ to the model without ξ (comparing Figure 3.11 vs. Figure 3.10) indicates the gain in power that results from including ξ in the

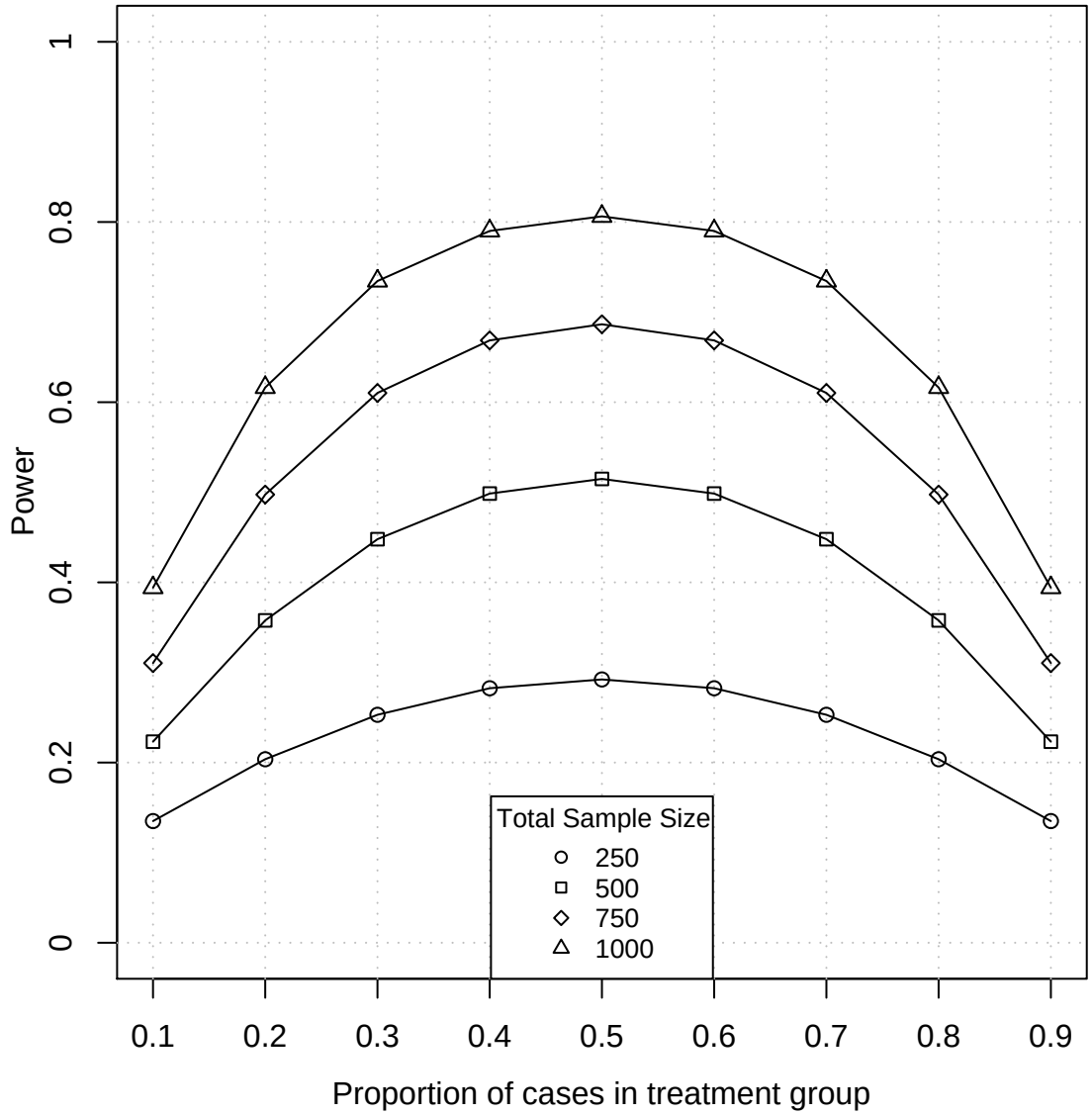


Figure 3.10: Power of the latent variable model (with or without ξ) to detect a small average treatment effect as a function of the proportion of cases in treatment group and total sample size given no interaction and ξ explaining 90% of outcome variance in the control group.

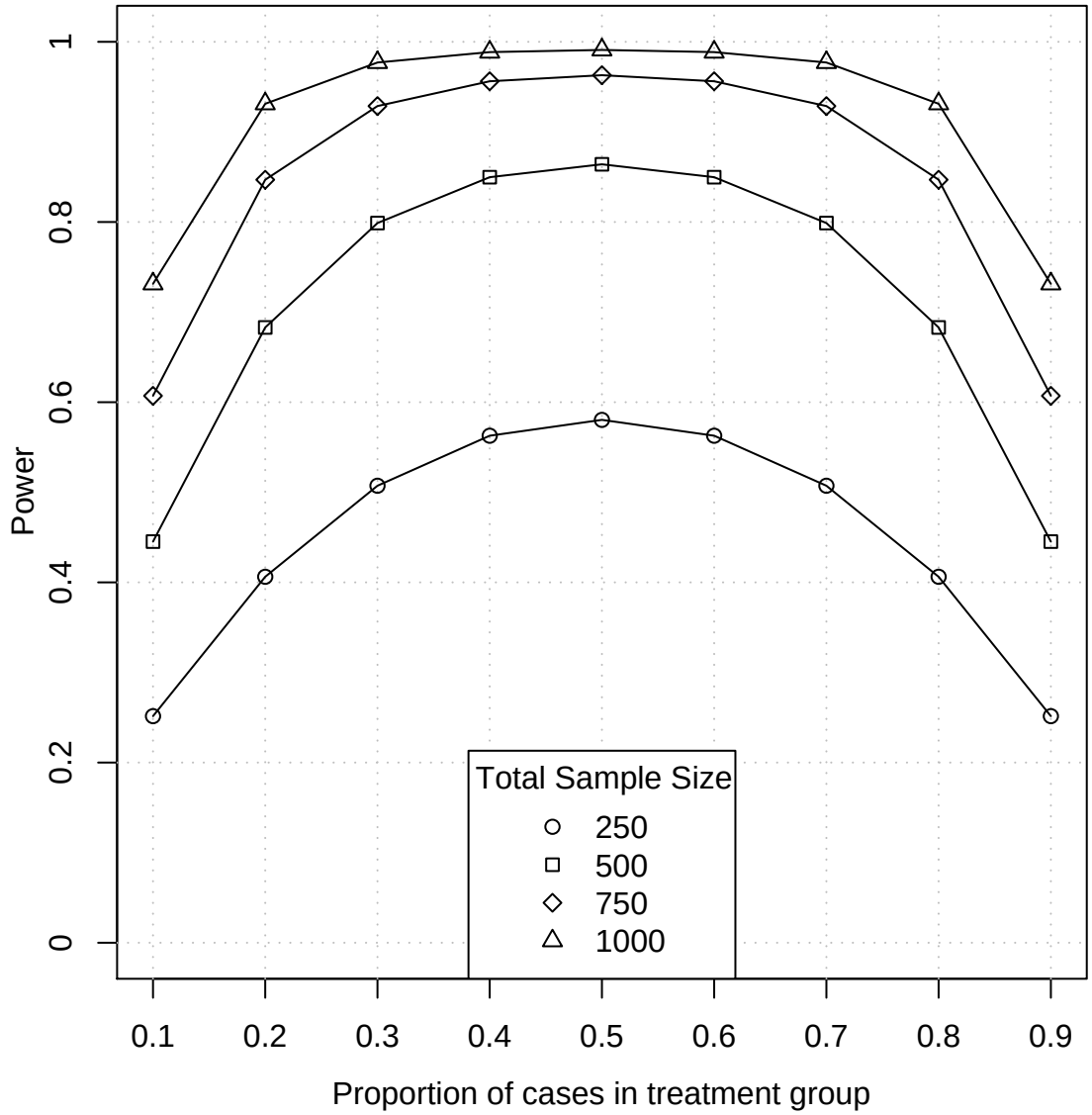


Figure 3.11: Power of the latent variable model including ξ to detect a small average treatment effect as a function of the proportion of cases in treatment group and total sample size given no interaction and ξ explaining 90% of outcome variance in the control group.

model. Figure 3.11 also shows symmetric power curves around the balanced case. The balanced case again yields the highest power and should be favored over the unbalanced design. It also should be mentioned that without an interaction effect present, the latent variable models including ξ share the same power whether they include the interaction or fix it to zero.

I now turn to the case with an interaction effect present. Figure 3.12 shows how the power varies as a function of the proportion of treatment-group observations for a given total sample size of 250, 500, 750, and $1,000$. The power of the latent variable model including ξ and the interaction is given for the case of a small average effect, a large interaction effect and ξ explaining a proportion of 90% of the latent outcome variance in group one.

The power curves are not symmetric around the balanced case where the proportion is .5. Choosing an unbalanced design in favor of more treatment observations is better than choosing an unbalanced design in favor of more control observations. This is because the outcome variance is larger in the treatment group than in the control group, whereas the reverse would hold if the treatment group variances were smaller. The reverse situation was verified by using an interaction effect of negative value that induced lower treatment group variance.

As mentioned before the average treatment effect may be estimated by a model that includes the latent covariate but fixes the interaction to zero even if an interaction is present. In the following, the latent model with the interaction fixed to zero is compared to the model that estimates the interaction with regard to the power to detect an average treatment effect.

Figure 3.13 shows the power of the latent variable model with an interaction effect fixed to zero for the same parameter values used in Figure 3.12 where an interaction effect is present and included in the model. Comparing

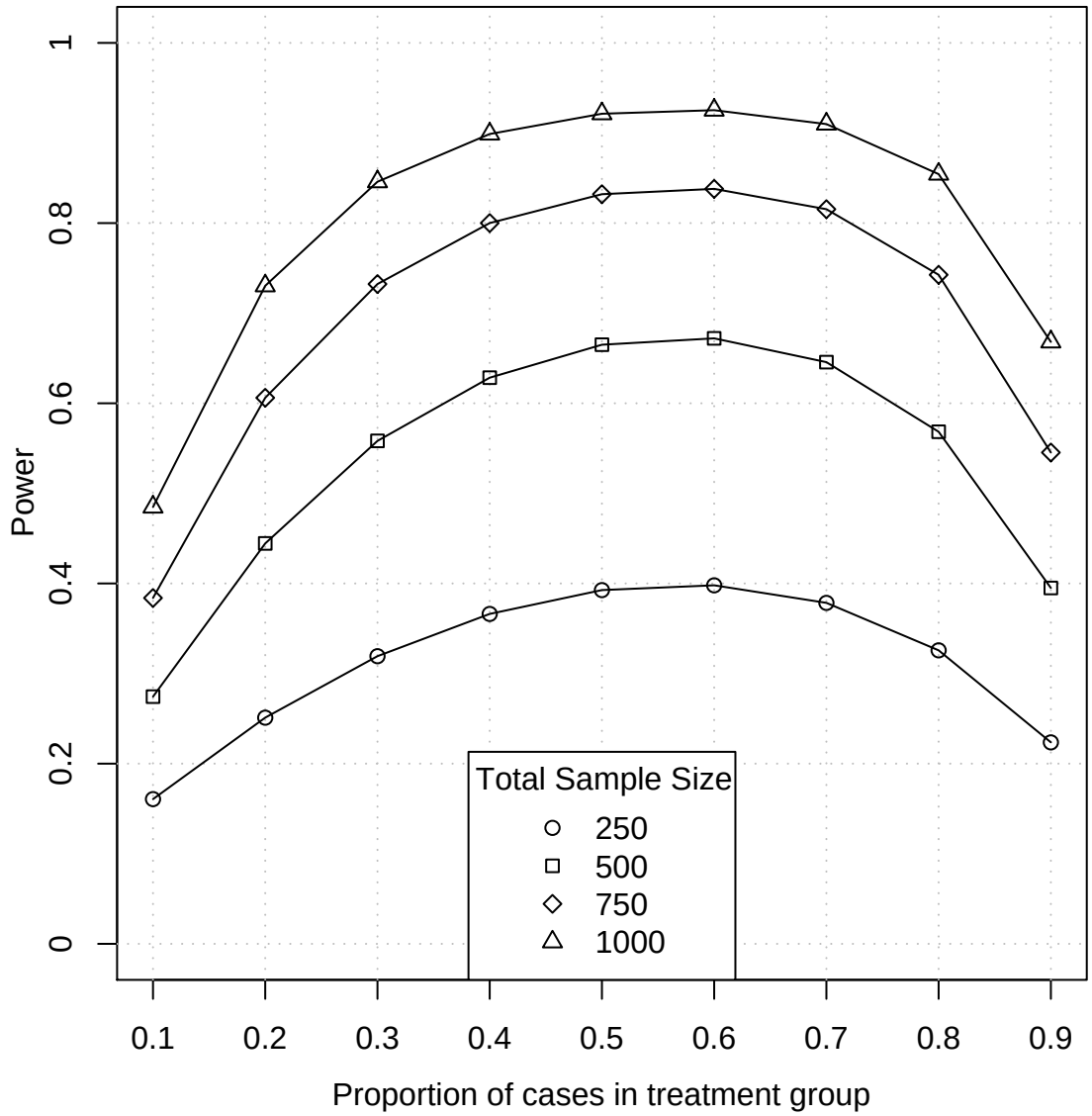


Figure 3.12: Power of the latent variable model including ξ and the interaction in order to detect a small average treatment effect as a function of the proportion of cases in treatment group and total sample size given a large interaction effect and ξ explaining 90% of outcome variance in the control group.

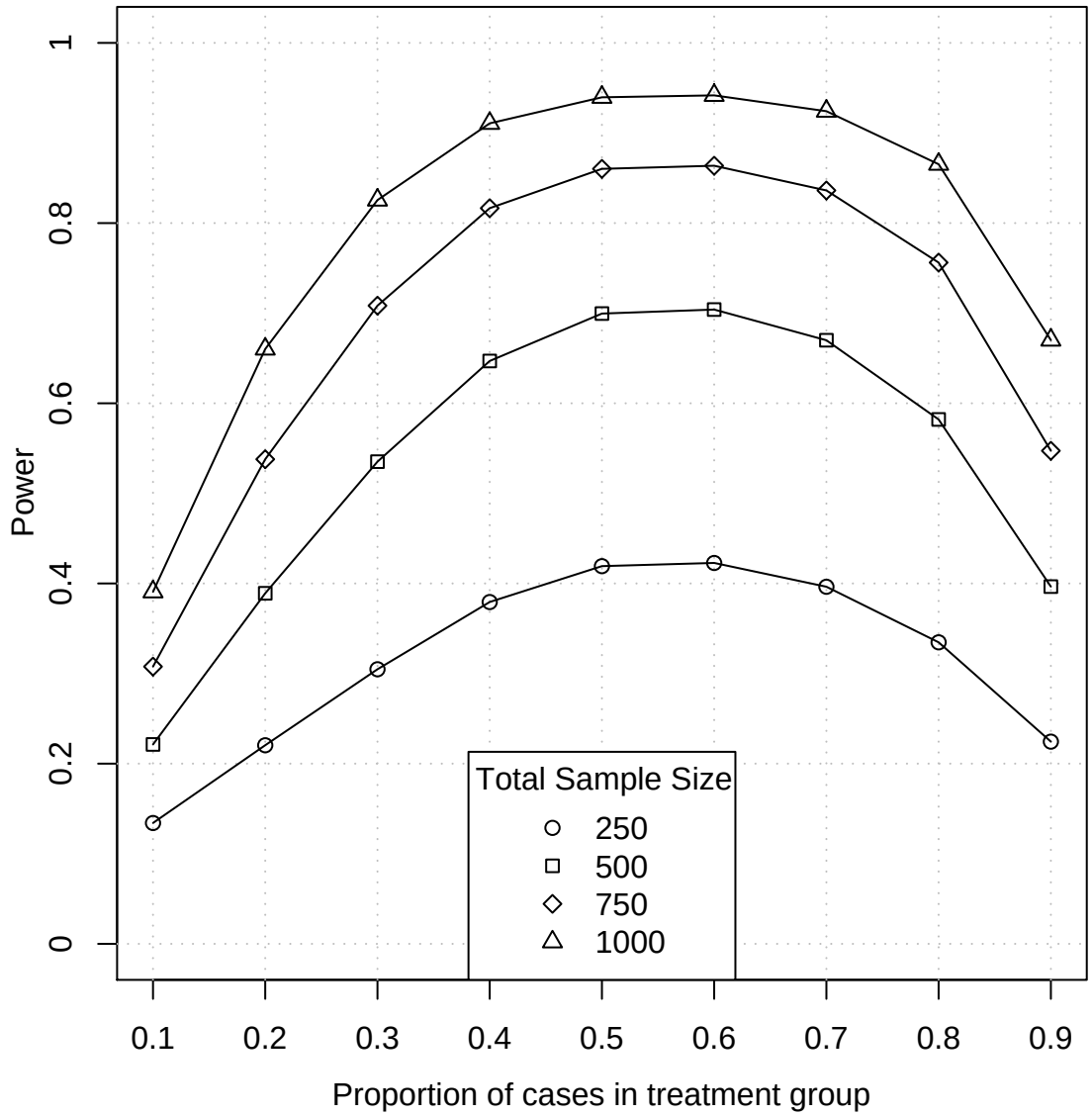


Figure 3.13: Power of the latent variable model fixing the interaction to zero in order to detect a small average treatment effect as a function of the proportion of cases in treatment group and total sample size given a large interaction and and ξ explaining 90% of outcome variance in the control group.

Figure 3.12 and 3.13 shows that the power curves are not identical. The model that ignores the interaction has slightly more power around the balanced case where the proportion is about .5 and the model that includes the interaction has slightly more power for small proportions of cases in the treatment group.

Figure 3.14 explicitly compares the two models for a total sample size of 750. It should be noted that the parameters for Figure 3.14 are chosen to show the largest power difference for the two compared models. The difference in power of the model including the interaction vs. the model without the interaction are smaller for all other sets of parameters used in this analysis. No differences are found for the case with no interaction present.

The differences of the compared models are small and may be considered negligible. A Monte Carlo study is conducted for the unbalanced design with 20% of cases in the treatment group, in order to verify the validity of these differences. The total sample size is 750 (150 observations in treatment group), the interaction parameter is set to a value of .8, and the average treatment effect is set to a value of .2. These parameters are chosen because the Satorra-Saris method yielded the largest advantage of the model including the interaction. The results are given in Table 3.5. All models estimate the average treatment effect without bias. The models differ in power. The models including ξ and the interaction show the highest detection rate of an average treatment effect indicating that the Satorra-Saris method correctly predicts the gain in power resulting from including the interaction in the model. Note however that the parameters for this example are chosen to resemble the most extreme case predicted by the Satorra-Saris method and that even for this example the difference is negligible to some extend.

We have seen that the Satorra-Saris method works very well to estimate

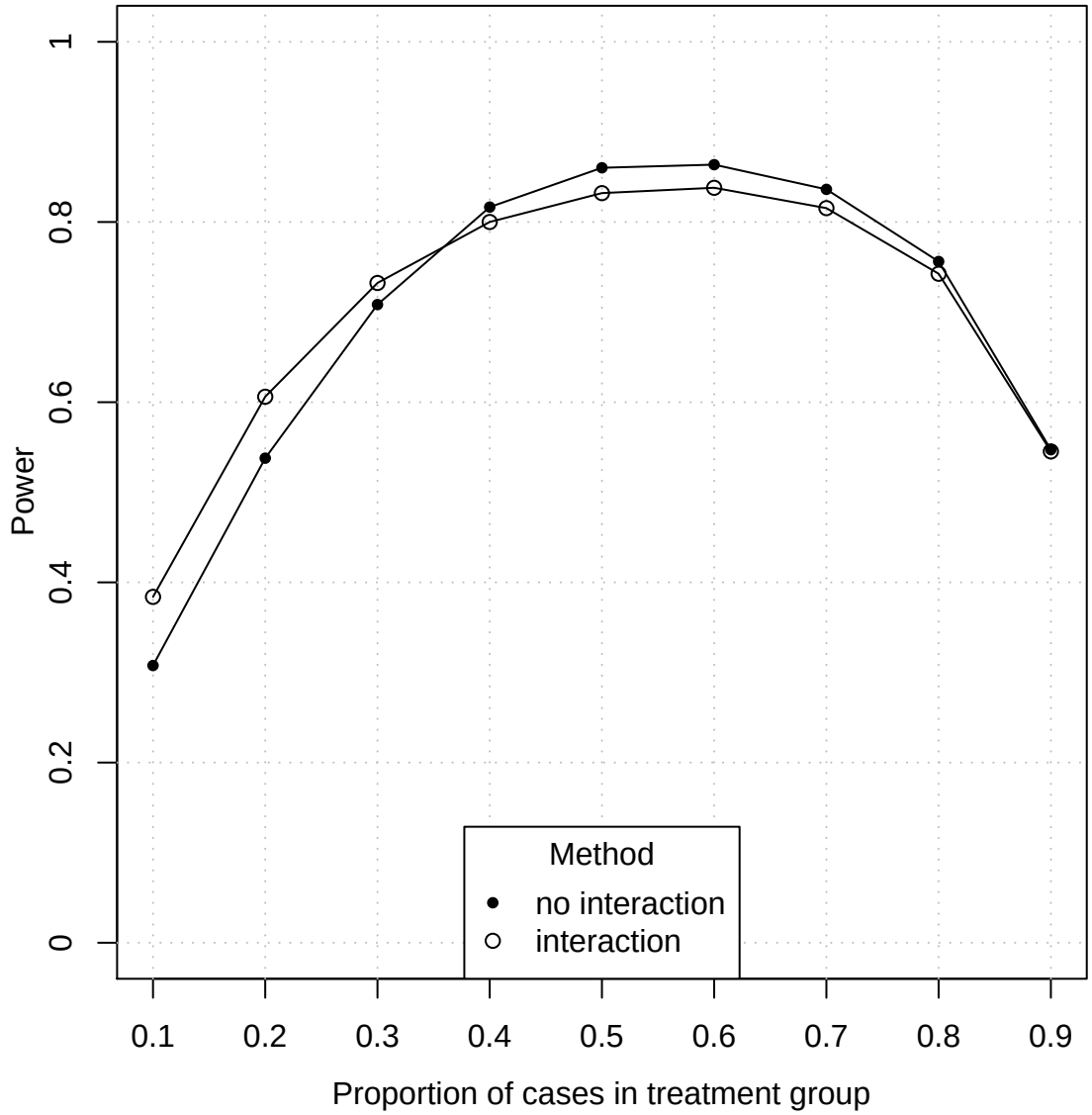


Figure 3.14: Power to detect an average treatment effect of the latent variable model including the interaction and the model without the interaction as a function of the proportion of cases in treatment group given a large interaction effect and a total sample size of 750.

Table 3.5

Results for the Monte Carlo Study for an Unbalanced Design (20% of Cases in Treatment Group) Given a Large Interaction and a Total Sample Size of 750.

Model	M	SD	rf_e	rf_o	C	W
1. w/ ξ, TC, w/ int.	0.196	0.088	0.606	0.617	1000	327
2. w/ ξ, TC, w/o int.	0.196	0.096	0.538	0.549	1000	128
3. w/ ξ, TE, w/ int.	0.196	0.088	0.606	0.617	1000	326
4. w/ ξ, TE, w/o int.	0.196	0.096	0.538	0.545	1000	130
5. w/o ξ, TC	0.208	0.163		0.247	927	941
6. w/o ξ, TE	0.194	0.156	0.245	0.241	1000	2
7. Mean Diff.	0.194	0.159		0.411	1000	0

Note. See Table 3.2 for a description.

power of the discussed latent multi-group model with regard to detect interaction effects as well as average treatment effects (for the considered examples). In summary, this chapter provides a description how to analyze average treatment effects based on models that include latent covariates and interactions between these latent covariates and the treatment. Standard SEM software may be used to implement the estimation and the test of the average treatment effect. It was also shown how to use Mplus in order to study power based on the Satorra-Saris method. The discussion was limited to randomized research designs. The following chapter will expand this topic to research designs that do not incorporate randomization.

Chapter 4

Non-randomized Designs

The last chapter described how multi-group structural equation modeling can be used to analyze average effects and interaction effects in the same modeling framework given that randomization is successfully implemented in the research design. In this chapter, several models are discussed and compared that may be used if randomization is not implemented in the design of a study. The key difference is that without randomization the expected values of the latent covariate might differ across groups. The multi-group approach described in the last chapter however, requires the assumption that these expected values are equal across groups. If this assumption does not hold, this chapter presents single-group as well as multi-group models that are applicable to analyze average effects. Monte Carlo studies are described that compare the performance of these approaches.

4.1 A Multi-group Approach to Analyze Non-randomized Designs

The average treatment effect was specified in Equation 3.7 for the model involving an interaction between a treatment and a latent covariate. Applying the multi-group model yielded the identical specification with the multi-group parameters given in Equation 3.22, which is repeated here

$$AE_{2-1} = \alpha^{(2)} - \alpha^{(1)} + \left(\beta^{(2)} - \beta^{(1)}\right) E(\xi). \tag{4.1}$$

It was mentioned that $E(\xi)$ is not a parameter of the multi-group model. The solution for a randomized design is to set the group specific mean equal in order to estimate the grand mean of the latent covariate. This is of course not feasible if the group means differ. Considering the following property

$$E(\xi) = E(\xi \mid X{=}1)\ P(X{=}1) + E(\xi \mid X{=}2)\ P(X{=}2), \tag{4.2}$$

shows, that Equation 4.1 can be written by replacing $E(\xi)$ with the two group specific means of ξ, which *are* parameters in the multi-group model. The average effect is then specified as

$$\begin{aligned} AE_{2-1} &= \alpha^{(2)} - \alpha^{(1)} \\ &\quad + \left(\beta^{(2)} - \beta^{(1)}\right) \Big(E(\xi \mid X{=}1)\ P(X{=}1) + E(\xi \mid X{=}2)\ P(X{=}2) \Big). \end{aligned} \tag{4.3}$$

Two cases have to be distinguished. First, the treatment probabilities $P(X{=}1)$ and $P(X{=}2)$ are fixed terms. This is the case, for example, if they are fixed by the research design. Second, these terms are observed as they (randomly) occur. In the first case, the values for the two terms may be treated as fixed numbers whereas in the second case they are parameters of the model, that are estimated by the observed relative group sizes.

Currently, standard structural equation software does not allow to include these treatment probabilities as parameters in a multi-group model. The observed values have to be written in the model and are always treated as fixed numbers. Consequently, no standard errors and covariances with other parameters are obtained. Using these values in constraints to test average effects may lead to problems (see, e. g., Nagengast, 2006). Monte Carlo Studies are described later that test the performance of the multi-group model (see Listing 4.4 on page 125 for a Mplus input example).

4.2 A Single-group Approach to Analyze Non-randomized Designs

So far the interaction between the treatment variable and the latent covariate has been treated with conventional structural equation modeling using multiple-group analysis, where the treatment variable (an observed unordered or categorical variable) represents the groups. It was shown in the last section, that it is possible to use multiple-group analysis to test interaction as well as average (or main) effects simultaneously. This was considered impossible beforehand (see, e. g Jaccard & Wan, 1996, p. 41). However as stated in the last section, there might be problems if the relative groups sizes are observed as they (randomly) occur.

The solution to this problem described in this section is to use a single-group approach that models the interaction between the treatment variable and the latent covariate. This type of interaction cannot be handled by conventional SEM. Special interaction modeling involving latent variables is needed, for example using the Joreskog-Yang approach (Jöreskog & Yang, 1996), 2SLS (Bollen, 1996), or the full-information maximum-likelihood ap-

proach of Klein and Moosbrugger (2000).

The interaction needed for the single-group model fits into the Mplus latent variable framework so that full-information maximum-likelihood estimation is possible (Asparouhov & Muthén, 2002). Klein and Moosbrugger (2000) pointed out the important efficiency and power advantages for interaction modeling by use of full-information maximum-likelihood estimation as compared to limited-information estimators such as the Joreskog-Yang and 2SLS approaches.

Consider again the interaction between the treatment and the latent covariate given in Equation 3.1, which is repeated here[1]

$$E(\eta \,|\, X, \xi) = \alpha + \beta_1 I_{X=2} + \beta_2 \xi + \beta_3 I_{X=2} \xi. \tag{4.4}$$

The Mplus approach to handling the interaction in Equation 4.4 uses a random slope variable (B. O. Muthén & Asparouhow, n.d.). The regression in Equation 4.4 can be written using two equations involving a random slope variable r and the residual ζ

$$\eta = \alpha + \beta_1 I_{X=2} + \beta_2 \xi + r I_{X=2} + \zeta \tag{4.5}$$

$$r = 0 + \beta_3 \xi + 0. \tag{4.6}$$

In Equation 4.5, a random slope is defined for the treatment indicator $I_{X=2}$. In Equation 4.6, the random slope is taken to be the same as the latent covariate ξ, except for a regression slope β_3, the interaction coefficient. In Equation 4.6, r is a latent variable that only contributes a single additional parameter. The corresponding intercept and the residual are set to zero. The two equations 4.5 and 4.6 indicate that the model is specified as a general

[1]Note that unlike conventional SEM (no interaction), the variance of η conditional on ξ changes as a function of ξ.

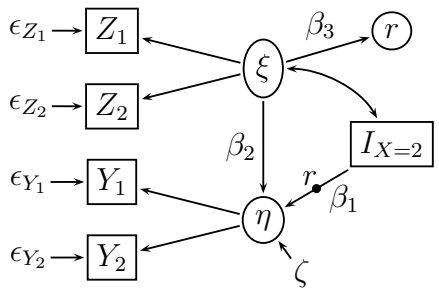


Figure 4.1: The SEM model extended to random slopes implemented in Mplus in order to include the interaction between the latent covariate and the treatment variable. The random slope variable, symbolized with the dot on the path from $I_{X=2}$ to η, constitutes a latent variable that only contributes the interaction parameter β_3 to the model.

structural equation model extended to random slopes. Such random slopes models can be handled in Mplus and will be used in the following to estimate average treatment effects.

The model described in Equation 3.1 to 3.4 fits in this extended SEM model. Figure 4.1 illustrates how this particular model is implemented in Mplus (compare to Figure 3.1 and 3.2). It is clear that r is a latent variable that only contributes the interaction parameter β_3 to the model. The arched double arrow between ξ and $I_{X=2}$ describes the correlation between the latent covariate and the treatment indicator. Given a non-randomized design this correlation can not be assumed to be zero (and the group means of ξ to equal).

At this point I want to mention that the random slope approach as it is described here assumes the same variances of the structural residual across groups. The previously described multiple group approach can model group

specific variances of the structural residual. Because the focus of this dissertation is on the average treatment effect, different error variance of the structural residual are no considered here. However there are theories that might imply different variances of the structural residual (see, e. g. Steyer, 2005) and it might be possible to extend the outlined random slope approach to incorporate these differences.

In the following, this random slope approach is used to analyze the average treatment effect. Again two ways of scaling the latent variables ξ and η will be discussed. Just like in the section describing the multiple-group analysis we will see that the default parameter settings of Mplus simplify the analysis of the average treatment effect.

4.2.1 Testing the Average Treatment Effect

One way to scale the latent variables ξ and η is to set the measurement slope of the first observed variable for each latent variable to one (i. e., $\lambda_{Z_1} = \lambda_{Y_1} = 1$). The first way described here completes the scaling of ξ by setting its mean to zero

$$E(\xi) = 0. \tag{4.7}$$

The scaling of η is completed by setting the structural intercept of η to zero

$$\alpha = 0. \tag{4.8}$$

For the given model the average treatment effect was given in Equation 3.7 on page 55. It is repeated here

$$AE_{2-1} = \beta_1 + \beta_3 \, E(\xi). \tag{4.9}$$

If ξ is scaled by setting $E(\xi) = 0$, then the average treatment effect is identified by β_1, the first order effect of the indicator variable. Hence,

```
DATA: FILE = data.dat;
VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
 USEVARIABLES = Y1 Y2 Z1 Z2 IX2;
 DEFINE: IX2 = 0; IF(X EQ 2) THEN IX2 = 1;
ANALYSIS: TYPE = RANDOM;
 ALGORITHM = INTEGRATION;
MODEL: ETA BY Y1-Y2;
 XI BY Z1-Z2;
 IX2 WITH XI;
 ETA ON XI;
 ETA ON IX2;
 R | ETA ON IX2;
 R ON XI;
 R@0;
 [R@0];
```

Listing 4.1: Mplus input to model the interaction between ξ (`XI`) and $I_{X=2}$ (`IX2`) using a random slope variable r (`R`).

estimating the average treatment effect is simply done by estimating β_1. No further specification of additional parameters is necessary. The standard error estimate of β_1 can be used to apply t-testing in order to test the average treatment effect against a hypothetical value. It is not necessary to manually compute the standard error estimate via the delta method.

Mplus offers two ways to implement this procedure. One way is to explicitly specify the random slope variable. Listing 4.1 gives the corresponding Mplus input for the given example. This method is discussed first because it shows how the random slope approach is implemented. The second way is more intuitive because the interaction is specified directly with an `XWITH` statement. Listing 4.2 gives the corresponding input for the given example and is discussed after the random slope specification.

```
1  DATA: FILE = data.dat;
2  VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
3   USEVARIABLES = Y1 Y2 Z1 Z2 IX2;
4   DEFINE: IX2 = 0; IF(X EQ 2) THEN IX2 = 1;
5  ANALYSIS: TYPE = RANDOM;
6   ALGORITHM = INTEGRATION;
7  MODEL: ETA BY Y1-Y2;
8   XI BY Z1-Z2;
9   IX2 WITH XI;
10  R | XI XWITH IX2;
11  ETA ON IX2 XI R;
```

Listing 4.2: Mplus input to model the interaction between ξ (`XI`) and $I_{X=2}$ (`IX2`) using the `XWITH` option. This input is equivalent to the one using the random slope specification given in Listing 4.1.

By selecting the analysis type as `RANDOM`, a model with random slopes will be estimated. The default estimator (`MLR`), a maximum likelihood estimator with robust standard errors using a numerical integration algorithm will be used. The standard errors are robust to non-normality and are computed using a sandwich estimator (L. Muthén & Muthén, 2004, p. 432). The computations require numerical integration which becomes increasingly more computationally demanding as the number of variables and sample size increase. The default integration uses rectangular (trapezoid) numerical integration with a default of 15 integration points per dimension. The model estimation uses full-information maximum-likelihood (L. Muthén & Muthén, 2004; see also Asparouhov & Muthén, 2002; Klein & Moosbrugger, 2000).

Similar to the multi-group model, the two latent variables are defined by lines 7 and 8. The measurement slopes of Y_1 and Z_1 are set to one per default. The mean of ξ is set to zero per default. Line 9 sets the correlation

between ξ and $I_{X=2}$ free, this line is required because the default is to set this correlation to zero. Line 10 and 11 specify that η is regressed on ξ and $I_{X=2}$ with the intercept set to zero per default. The remaining lines implement the interaction in this regression. Line 12 specifies r as a random slope variable in the regression of η on $I_{X=2}$. Line 13 to 15 specify this random slope variable to be the same as ξ, except for a scaling factor, the interaction coefficient β_3. Line 13 specifies this regression that contributes β_3 to the model, line 14 and 15 specify the intercept and the residual of this regression to be zero respectively.

Mplus also offers a more intuitive way to specify this interaction. Listing 4.2 gives the input corresponding to Listing 4.1. Line 10 of Listing 4.2 defines the "interaction variable" `R`, which then occurs in line 11 (besides `IX2` and `XI`) as a regressor on which η is regressed on. Both specifications are treated to be identical throughout the following discussion of the model and the Monte Carlo studies[2].

The summary of the analysis results given in the Mplus output contain the log-likelihood for the analysis model but no chi-square statistic. Model difference testing is possible using the log-likelihood based on a method developed by Satorra (2000). It is however not needed here. The interested reader will find more information on the Mplus website under *Difference Testing Using the Loglikelihood* (n.d.).

The parameter estimation and test for the interaction and the average treatment effect do not require model difference testing and are readily available from the output under the model results. The parameters of special

[2]The two methods differed for a few data sets of the Monte Carlo studies below with regard to their default starting values. The `XWITH` statement produced slightly less warnings. The performance of both methods with regard to parameter estimation and overall model estimation was almost identical.

interest here include β_3, the interaction, which is found in the row labeled `R ON XI`, as well as β_1, the average treatment effect, which is found in the row labeled `ETA on IX2`. The standard errors (labeled `S.E.`) produced during model estimation are determined by the MLR estimator described above. The values labeled `Est./S.E.` contain the values of the parameter estimate divided by the standard error (column 1 divided by column 2). This statistical test is an approximately normally distributed quantity (z-score) in large samples. The critical value for a two-tailed test at the .05 level is an absolute value greater than 1.96.

We have seen how Mplus can be used to analyze the interaction between a treatment and a latent variable as well as the average treatment effect. The described model uses a structural equation modeling framework extended to random slopes. The single-group model is estimated using full-information maximum-likelihood and avoids potential problems that may be involved in the corresponding multi-group approach mentioned in section 4.1.

4.3 Monte Carlo Studies

In this section, Monte Carlo Studies are conducted in order to assess the performance of the single-group approach described in the previous section as well as the multi-group approach described in section 4.1 with regard to estimating and testing the average treatment effect. It was argued that the multi-group approach is more flexible than the single-group approach, because it allows different variances of the structural residual across groups. However, it was also argued that the multi-group approach does not treat all model parameters as such and problems might arise if the treatment probability $P(X{=}2)$ is estimated from the sample.

The single-group approach on the other side avoids this problem because it allows to model the average treatment effect in a sound manner. It is however much more computationally demanding than the multi-group approach and currently only implemented in Mplus. The following Monte Carlo studies are conducted to address theoretical concerns and to give advice to practitioners. The comparison focuses on the biasedness of the estimators of the average treatment effect as well as the biasedness of the standard error estimators of the average treatment effect. Because the data generation for these Monte Carlo studies is more complex than in the Monte Carlo studies for randomized designs it is discussed in detail.

4.3.1 Data Generation

In the Monte Carlo studies for randomized designs Mplus was used to generate and analyze the data (see Chapter 3.5). In the following Monte Carlo studies, the data is generated with the statistical programming environment R (R Development Core Team, 2006) and analyzed with Mplus. The data is generated in order to simulate a non-randomized design where the treatment assignment is expected to depend on other variables. Here, the treatment assignment depends only on the latent covariate ξ. Hence, including the latent covariate in the analysis yields causally unbiased estimates for the conditional and average treatment effects (see Steyer et al., 2007, for a detailed discussion on the topic on causality).

A logistic regressive dependency is chosen in order to model the dependency of the assignment probabilities on the latent covariate. The program to generate data is given in Listing B.2 on page 149. It is written in order to simulate the following random experiment: From a population of units draw one unit, record the values of two observed variables Z_1 and Z_2 that measure

the latent covariate ξ. Randomly assign the unit to one of the treatment groups. This random assignment is *not* completely randomized. Instead, the assignment probabilities for each group depend on the latent covariate and are determined by logistic functions

$$P(X{=}2 \,|\, \xi) = 1/\left(1 + \exp(l_0 + l_1\xi)\right) \tag{4.10}$$

$$P(X{=}1 \,|\, \xi) = 1 - P(X{=}2 \,|\, \xi). \tag{4.11}$$

The treatment is applied and the values of the two observed outcome variables Y_1 and Y_2 that measure the latent outcome variable η are recorded. The value of the latent outcome variable η depends on the treatment and the latent covariate based on the following equation

$$\eta = \alpha + \beta_1 I_{X=2} + \beta_2\xi + \beta_3 I_{X=2}\xi + \zeta. \tag{4.12}$$

The values for ζ, the residual of this regression, are taken from a normally distributed random variable with expected value zero and variance $Var(\zeta) = .5$.

The four observed variables are computed based on the measurement models described in Equations 3.3 to 3.4 involving normally distributed measurement residuals, ϵ_{Z_1}, ϵ_{Z_2}, ϵ_{Y_1}, and ϵ_{Y_2} each with variance .5. All residuals are generated independent from each other and the latent variables.

The difference to the Monte Carlo Studies in Chapter 3.3 is that the treatment assignment is not (completely) randomized. Instead, the treatment assignment is conditionally randomized with the assignment probabilities depending on the latent covariate. In the current version of Mplus it is not possible to generate such data internally (Linda Muthén, personal communication, May 16, 2006). The data is generated externally using R. Listing B.2 on page 149 gives the corresponding code.

This data generation proceeds in several steps (all parameters given in parentheses refer to the program code in Listing B.2). The values for the latent covariate ξ (`xi`) are generated by drawing `N` units from a standard normal distribution (i. e. $E(\xi) = 0$ and $Var(\xi) = 1$).

The values of each observation on the latent covariate ξ are then used to determine the assignment probabilities for the treatment groups. The values for the treatments are taken from a random variable with a discrete probability distribution, which takes value 2 with conditional probability $P(X{=}2 \mid \xi)$ and value 1 with conditional probability $1 - P(X{=}2 \mid \xi)$. The values of these conditional assignment probabilities are determined by the logistic functions given in Equations 4.10 and 4.11.

The parameters of these functions, l_0 (`l0`) and l_1 (`l1`), are varied in order to represent different situations. The threshold parameter l_0 (`l0`) varies in order to change the (unconditional) assignment probabilities $P(X{=}1)$ and $P(X{=}2)$, in other words the expected relative group sizes. The slope parameter l_1 (`l1`) varies in order to change the dependency of the treatment assignment on the latent covariate.

The values of η are calculated based on the regression given in Equation 4.12. The structural intercept α as well as the first order effect of the treatment indicator β_1 are fixed to zero. This is done to insure that the hypothesis of no average treatment effect in the population is true. Given $E(\xi) = 0$ and $\beta_1 = 0$, the average treatment effect is zero (see Equation 4.9 on page 110).

The first order effect of ξ is set to $\beta_2 = \sqrt{.5}$ in order to achieve a moderate dependency of η on ξ given that no interaction effect is present. The interaction effect β_3 is set by choosing a value for `b3`. The structural part of the SEM model is completed by adding the structural residual ζ, a normally dis-

tributed random variable with expected value zero and variance $Var(\zeta) = .5$. With no interaction effect present (i. e. $\beta_3 = 0$), the latent covariate can be expected to explain 50% of the variance of η.

The values for the observed variables Z_1, Z_2, Y_1, and Y_2 measuring the latent covariate and the latent outcome respectively are computed according to the measurement models given in Equations 3.3 and 3.4. For simplicity the parameters[3] are set to the (constant) values:

$$\nu_{Z_k} = \nu_{Y_k} = 0 \tag{4.13}$$

$$\lambda_{Z_k} = \lambda_{Y_k} = 1 \tag{4.14}$$

$$Var(\epsilon_{Z_k}) = Var(\epsilon_{Y_k}) = 0.5, \tag{4.15}$$

where $k = 1, 2$. The values for the measurement residuals ϵ_{Z_1}, ϵ_{Z_2}, ϵ_{Y_1}, and ϵ_{Y_2} are taken from independently normally distributed random variables each with expected value zero and variance .5.

The consequences for the distributions of the involved variables are as follows. Although ξ is normally distributed, the distribution of ξ conditional on X (i. e., per group) deviates from normality due to the logistic regressive dependency of X on ξ. Hence, the distributions of Z_1 and Z_2 conditional on X are non-normal. The distribution of η deviates from normality due to the interaction. The distribution of η conditional on X deviates from normality due to the non-normality of the distribution of ξ conditional on X. Hence, the distribution of Z_1 and Z_2 as well as the conditional distribution of Y_1 and Y_2 given treatment are non-normal.

Despite these deviations from non-normality the following assumptions

[3]The parameters of the measurement models are chosen to be identical to the parameters used in the Monte Carlo studies for randomized designs described in Chapter 3.3 (see Equation 3.33). The reliabilities differ to some extend from the reliabilities given in Chapter 3.3 due to the dependency of X on ξ.

are met. The structural residual ζ is normally distributed and independent from ξ. Both properties also hold conditional on X. The distributions of Z_1 and Z_2 as well as the distribution of Z_1 and Z_2 conditional on X do not depend on η, ζ, Y_1, or Y_2. Given these assumptions, minimizing the likelihood function of the general latent variable framework for the multiple group analysis given in Equation 3.14 leads to maximum likelihood estimation regardless of described deviations from normality (Bollen, 1989, 126–127; Johnston, 1984, 281–285; Jöreskog, 1973, 94).

Two Monte Carlo studies are conducted. The first one considers the case where the two (unconditional) assignment probabilities $P(X{=}1)$ and $P(X{=}2)$ are equal. The threshold parameter of the logistic regression of X on ξ is set to $l_0 = 0$ (balanced case). In the second study, the assignment probabilities are unequal. The probability of being assigned to the control group is about twice the probability of being assigned to the treatment. In both Monte Carlo studies the data was analyzed with three methods:

- Mplus approach to handling interactions by extending SEM with a random slope variable; referred to as the single-group approach (see section 4.2)

- the multi-group-approach with the assignment probability (written in the constraint for the average treatment) fixed to the *true population* value (see section 4.1, and Equation 4.16 below)

- the multi-group-approach with the assignment probability *estimated* from each replication

4.3.2 Monte Carlo Study with Equal Group Sizes

In the first Monte Carlo study, the threshold of the logistic regression of X on ξ, is set to $\texttt{10} = 0$ and held constant. Hence, the (unconditional) assignment probabilities are $P(X{=}1) = P(X{=}2) = .5$. Note however, that the actual relative group sizes vary from sample to sample.

Design

The performance of the average treatment effect estimators are studied by systematically varying the following parameters[4] in a fully crossed factorial design: *total sample size* (`N`), *dependency of* X *on* ξ (`l1`), and the *interaction* (`b3`).

The **total sample size** is either set to $\texttt{N} = 400$ or 1000. These values are chosen based on the discussion of the power analyzes in chapters 3.5 to 3.5.3. As shown, a sample size of 400 will certainly not be sufficient to detect a small interaction effect and a small average treatment effect with a desirable power of about .8. Nevertheless, such a sample size is often used in behavioral studies and is therefore included in the study.

The **dependency** of X on ξ is varied by choosing two different values for the slope parameter $l_1 = -1$ and $l_1 = -5$. The corresponding expected dependencies of X on ξ are expressed in terms of the correlation between $I_{X=2}$ and ξ. For $l_1 = -1$ numerical integration yields an expected correlation of $Corr(\xi, I_{X=2}) \approx 0.413$ and is referred to as a *moderate* dependency of the treatment on the latent covariate. For $l_1 = -5$ the expected correlation between the latent covariate and the treatment indicator is $Corr(\xi, I_{X=2}) \approx 0.751$ and is referred to as a *strong* dependency of the treatment on the latent covariate.

[4]The parameters in the parentheses refer to the R input given in Listing B.2.

The **interaction** between X and ξ is varied between the following values $\beta_3 = 0$, .2, .5, .8, 1.5, 3, and 10. The interpretation of these values calls for measures of effects size. The classification used here follows the outline in chapter 3.5. Throughout chapter 3.5 the effect size of the interaction is determined by dividing the interaction parameter by $Var(\eta \mid X{=}1)$, the variance of η in the control group. All parameters in the power analysis for randomized designs are chosen so that $Var(\eta \mid X{=}1) = 1$, yielding equivalence of interaction parameter and corresponding effect size.

Given the data generation for non-randomized designs, $Var(\eta \mid X{=}1)$ depends on the regression coefficients and additionally on the parameters of the logistic regression of X on ξ. Hence, changing the parameters of the logistic regression would change the effect size of the interaction.

The effect size of the interaction would have to be adjusted according to the dependency of X on ξ. Given the moderate dependency of X on ξ (i. e. `11=-1`) the variance of η in group one is approximately .91. Hence, the effect size of the interaction would be about 10% larger than the corresponding interaction parameter. Given the strong dependency of X on ξ (i. e. `11=-5`) the variance of η in group one is approximately .72. Hence, the effect size of the interaction would be 30% larger than the corresponding interaction parameter.

To avoid this complication, the measures of effect size used for the interaction effect are determined by ignoring the dependency of X on ξ. Given no dependency of X and ξ the variance of η in group one remains one (i. e., $Var(\xi \mid X{=}1) = 1$) thus yielding equivalence of interaction parameter and corresponding effect size. Consequently, the interpretation of the effect sizes of the interaction parameter is invariant to changes of the dependency of X on ξ and is comparable to the effect sizes given in chapter 3.5.

Following Cohen's (1988) classification, the values of the interaction parameter $\beta_3 = .2, .5, .8$ are referred to as *small*, *medium*, and *large* interaction effects respectively. The values $\beta_3 = 1.5$, and 3 are both referred to as *very large* interaction effects. The value $\beta_3 = 10$ will seldom occur in real data and is included here merely for theoretical purposes.

4.3.3 Data analysis

The data are analyzed with the three methods listed above. First, the single-group approach using the Mplus SEM model extended to random slopes is applied using the code given in Listing 4.3. This method will be referred to as ***single-group approach***. The code is almost the same as the one given in Listing 4.2. The only difference between Listing 4.3 and Listing 4.2 occur in line 1 and 2, stating that the data to be analyzed are of type `MONTECARLO`, meaning that many data sets (replications) are to be analyzed with the corresponding file names listed in `mcreplist.dat`.

The outputs of the code in Listing 4.3 include a summary of the overall model fit containing mean and standard deviation of the log-likelihood test statistic over the replications of the Monte Carlo analysis. The summary of results includes the true population value for each parameter, the average of the parameter estimates across replications, the standard deviation of the parameter estimates across replications, the average of the estimated standard errors across replications, the means square error for each parameter (M.S.E.), 95% percent coverage, and the proportion of replications for which the null hypothesis that a parameter is equal to zero is rejected at the .05 level.

The main focus of this study is on the average treatment effect. The default scaling of the latent variables is the same as described in section 4.2.1,

```
DATA: FILE = mcreplist.dat;
 TYPE = MONTECARLO;
VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
 USEVARIABLES = Y1 Y2 Z1 Z2 IX2;
 DEFINE: IX2 = 0; IF(X EQ 2) THEN IX2 = 1;
ANALYSIS: TYPE = RANDOM;
 ALGORITHM = INTEGRATION;
MODEL: ETA BY Y1-Y2;
 XI BY Z1-Z2;
 IX2 WITH XI;
 R | XI XWITH IX2;
 ETA ON IX2 XI R;
```

Listing 4.3: Mplus input for the Monte Carlo study using the single-group approach in order to estimate and test the average treatment effect.

where it was shown that the average treatment effect is identified by β_1, the first order effect of the indicator for treatment two. This parameter is labeled `ETA ON IX2`. The average of the parameter estimates across replications (value of column two `ESTIMATES Average`) is used to assess the bias (see below). Column three, labeled `Std. Dev.`, gives the standard deviation of the parameter estimates across the replications. Because the number of replication is 1000, this value is considered to be the population standard error. Column four, labeled `S.E. Average`, gives the average of the estimated standard errors across replications. These two values are used to assess the standard error bias (see below).

The data are also analyzed with two versions of the ***multi-group approach*** described in chapter 4.1. In both versions the latent variables are scaled so that the mean of ξ in group one and the structural intercept of η in group one ($\alpha^{(1)}$) are both zero. Hence, the specification of the average

treatment effect (Equation 4.3) reduces to

$$AE_{2-1} = \alpha^{(2)} + \left(\beta^{(2)} - \beta^{(1)}\right) E(\xi \mid X{=}2)\ P(X{=}2). \tag{4.16}$$

The first version uses $P(X{=}2)$, the assignment probability of the population, whereas the second version uses $\widehat{P}(X{=}2)$, the assignment probability estimated from the sample by the proportion of cases in group two. The first version is applicable if the (relative) group sizes are under control by the researcher. In this case, the quantity used in the constraint is fixed and does not change between replications.

The second case, however, is more often encountered in applied research studies. The (population) probabilities are not under control by the researcher and are not known. Hence, the assignment probability has to be estimated by the proportion of cases in group two which (randomly) occurs in the sample. Because Mplus (and other standard SEM) software does not allow to treat this quantity as a parameter in a multi-group model, the observed value is simply written in the model constraint. Consequently, no standard errors and covariances with other parameters are obtained. The Monte Carlo study investigates (for a few examples) potential shortcomings of this method.

The code for version one (using the population treatment probability) is given in Listing 4.4. Line 11 specifies the average treatment effect with a nonlinear constraint according to Equation 4.16. The `MONTECARLO` statement is used to obtain summary results over the replications.

The output of the code in Listing 4.4 includes a summary of the overall model fit containing mean and standard deviation of the chi-square test statistic over the replications of the Monte Carlo analysis. The summary of results includes information about the same measures as the output of Listing 4.3. The results for the average treatment effect are found under `New/Additional Parameters`.

```
1  DATA: FILE = mcreplist.dat; TYPE = MONTECARLO;
2  VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
3   GROUPING = X (1 = G1 2 = G2);
4  ANALYSIS: TYPE = MEANSTRUCTURE;
5  MODEL: XI BY Z1 Z2; ETA BY Y1 Y2; ETA ON XI;
6  MODEL G1: ETA ON XI (BE1);
7  MODEL G2: ETA ON XI (BE2);
8   [XI] (MXI2);
9   [ETA] (AL2);
10 MODEL CONSTRAINT: NEW(AE);
11  AE = AL2 + (BE2 - BE1) * MXI2 * 0.5;
12 OUTPUT: TECH9;
```

Listing 4.4: Mplus input for the Monte Carlo study using the multi-group approach with the population treatment probability in order to estimate and test the average treatment effect.

For the second version of the test, the `MONTECARLO` statement is not feasible because the values for the relative group sizes change between replications. It is therefore necessary to write an input file and save the corresponding results for every replication. Listing 4.5 provides an example for the version using the sample proportion of cases in group two. Line 11 specifies the average treatment effect with a non-linear constraint according to Equation 4.16. The results for the average treatment effect are again found under `New/Additional Parameters` labeled `AE`. The results are summarized over all replications using R.

4.3.4 Results

The performance of the three methods is evaluated with regard to the bias of the average effect estimator as well as the bias of the standard error estimator

```
DATA: FILE = mcrep1.dat;
VARIABLE: NAMES ARE Y1 Y2 Z1 Z2 G;
 GROUPING IS X (1 = G1 2 = G2);
 ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL: XI BY Z1 Z2; ETA BY Y1 Y2; ETA ON XI;
MODEL G1: ETA ON XI (BE1);
MODEL G2: ETA ON XI (BE2);
 [XI] (MXI2);
 [ETA] (AL2);
MODEL CONSTRAINT: NEW(AE);
 AE = AL2 + (BE2 - BE1) * MXI2 * 0.51;
OUTPUT: TECH1;
 SAVEDATA: RESULTS = mcrep1.res;
```

Listing 4.5: Mplus input for the Monte Carlo study using the multi-group approach with the relative group size from the sample in order to estimate and test the average treatment effect.

for the average effect. Assessing bias of a parameter estimator is usually done by subtracting the population value from the average over all estimates and dividing the result by the population value. Because the population value of the average effect is zero, bias is assessed simply by evaluating the average over all estimates. Table 4.1 gives the results for the three methods and all parameter combinations.

All three models yield averages that are sufficiently close to the true population value of zero. Hence, for the given parameter combinations, all three methods provide unbiased estimates of the average effect. The SEM approach extended to random slopes provides an unbiased estimate of the average treatment effect. Also, both multi-group models provide unbiased estimates, despite the deviations from normality of the distributions of several

variables. Ignoring the fact that the relative group size is a random variable does not seem to negatively effect the average effect estimation (version two of the multi-group test).

Standard error bias is assessed by subtracting the population standard error value (the standard deviation of all average effect estimators) from the average standard error value and dividing this number by the population standard error value and multiplying by 100. Table 4.2 gives the results for the three methods and all parameter combinations.

All three models yield negligible standard error biases for small, medium and large interaction effects. The standard error biases of the single-group model as well as the standard error biases of the multi-group model with the population treatment probability are acceptable for all parameter combinations (values less than five). Only the multi-group model using the estimated treatment probability yields inflated standard error biases for very large interaction effects ($\beta_3 = 1.5$, 3, and 10). The stronger the dependency of the treatment on the latent covariate the larger the inflation. The sample sizes does not seem to have an effect on this inflation.

4.3.5 Monte Carlo Study with Unequal Group Sizes

In the second Monte Carlo study the threshold parameter of Equation 4.10 is set to $l_0 = 1$ and the resulting assignment probabilities are $P(X{=}1) \approx 0.7$ and $P(X{=}2) \approx 0.3$. These values are computed by using numerical integration. Consequently, the expected proportion of cases in the treatment group is less than a third of the total sample size.

The performance of the average treatment effect estimators is studied with a smaller design than in the previous Monte Carlo study. The reason for this is the computational demand of the SEM approach extended to random slope

Table 4.1

Averages of the Average Effect Estimators Over All 1000 Replications for the Three Methods. Varied Parameters are: Interaction, *Correlation Between Treatment and* ξ *(*moderate *or* strong*), and Sample Size (*400 *or* 1000*).*

	Interaction						
	0	0.2	0.5	0.8	1.5	3	10
	Singlegroup model extended to random slopes						
moderate							
400	0.00	0.00	−0.01	−0.01	−0.00	0.01	0.01
1000	−0.00	−0.00	0.00	−0.00	−0.00	0.01	0.00
strong							
400	−0.00	−0.00	−0.00	−0.01	−0.00	0.00	−0.00
1000	−0.00	−0.00	−0.00	0.00	−0.01	0.00	0.01
	Multigroup model with population treatment probability						
moderate							
400	0.00	−0.00	−0.01	−0.01	−0.00	−0.01	−0.01
1000	−0.00	−0.00	0.00	−0.00	−0.00	0.00	−0.01
strong							
400	−0.01	−0.00	−0.01	−0.02	−0.01	−0.01	−0.02
1000	−0.00	−0.00	−0.00	−0.00	−0.01	−0.01	−0.01
	Multigroup model with estimated treatment probability						
moderate							
400	0.00	−0.00	−0.01	−0.01	−0.00	−0.01	−0.00
1000	−0.00	−0.00	0.00	−0.00	−0.00	−0.00	−0.00
strong							
400	−0.01	−0.00	−0.01	−0.02	−0.01	−0.01	−0.03
1000	−0.00	−0.00	−0.00	−0.00	−0.01	−0.01	−0.01

Table 4.2

Standard Error Biases of the Three Models. Varied Parameters Are: Interaction, *Correlation Between Treatment and* ξ (moderate *or* strong)*, and Sample Size (*400 *or* 1000*).*

	Interaction						
	0	0.2	0.5	0.8	1.5	3	10
	Singlegroup model extended to random slopes						
moderate							
400	−1.80	−3.17	−0.80	−3.52	−3.21	−1.07	1.44
1000	−1.01	−1.32	−2.17	−2.45	−0.72	−2.99	−0.31
strong							
400	−3.94	−3.99	3.58	−1.71	−1.34	−0.55	1.89
1000	−0.79	−1.86	−0.43	1.45	−1.85	−0.86	−0.91
	Multigroup model with population treatment probability						
moderate							
400	−2.22	−3.28	−1.10	−3.20	−2.75	−1.88	0.61
1000	−1.01	−1.48	−2.03	−2.34	−0.49	−1.16	−0.24
strong							
400	−4.33	−3.97	3.86	−0.19	−4.62	0.04	**5.75**
1000	−1.01	−2.29	0.11	0.10	−1.66	1.87	−0.97
	Multigroup model with estimated treatment probability						
moderate							
400	−2.23	−3.46	−1.62	**−5.22**	**−6.26**	**−6.02**	**−6.10**
1000	−0.90	−1.53	−2.95	−3.35	−3.79	**−7.69**	**−6.49**
strong							
400	−4.32	−3.87	2.87	−3.03	**−7.73**	**−9.80**	**−12.09**
1000	−1.00	−1.94	−1.25	−1.10	**−5.19**	**−8.83**	**−15.12**

Note. Values with an absolute value larger than five are printed bold and are considered severely biased.

due to numerical integration. The only factor included in the design of the Monte Carlo study here is the interaction effect between X and ξ. The total sample size is set to $\texttt{N} = 400$. The slope parameter is set to $l_1 = -1$. The corresponding dependency of X on ξ is $Corr(I_{X=2}) \approx 0.39$ and is considered to be *moderate*. The values of β_3 are systematically varied between the same values as in the previous study. The effect size of the interaction is again considered to be equivalent to the corresponding parameter values.[5]

Data are again analyzed with the three methods described in the previous section. The input for the single-group SEM approach extended to random slope is exactly the same (Listing 4.3). For the two multi-group versions the values of the assignment probability for treatment is set to either the true population value $P(X{=}2) \approx 0.303$ (line 11 in Listing 4.4) or the sample value (line 11 in Listing 4.5).

Results

The performance of the three methods is evaluated with regard to the bias of the average effect estimator as well as the bias of the standard error estimator for the average effect. Because the population value of the average effect is zero, bias is assessed by evaluating the average over all estimates. Table 4.3 gives the results for the three methods.

All three models yield averages that are sufficiently close to the true population value of zero. Hence, all methods provide unbiased estimates of the average effect for the given case of unequal treatment probabilities.

[5]The norm to determine effect size is again the variance of η in group one as it would be without a dependency of X on ξ (i. e., $Var(\eta \mid X{=}1) = 1$). The expected variance of η in group one for the given dependency of X on ξ is actually $Var(\eta \mid X{=}1) = .93$.

Table 4.3

Estimation Biases of the Second MC Study for the Models Including the Interaction Given a ***Moderate*** *Treatment Latent Covariate Correlation.*

Interaction						
0	0.2	0.5	0.8	1.5	3	10
Singlegroup model extended to random slopes						
0.00	0.01	0.00	0.01	0.01	0.03	0.01
Multigroup model with population treatment probability						
−0.00	0.00	−0.00	0.00	−0.01	−0.00	−0.05
Multigroup model with estimated treatment probability						
−0.00	0.00	−0.00	0.00	−0.01	−0.00	−0.04

The SEM approach extended to random slopes gives an unbiased estimate of the average treatment effect. It performs well for the given examples with regard to estimating and testing the average treatment effect. Both multigroup models provide unbiased estimates. Note again that the distribution of several variables in the given examples deviate from normality. Ignoring the fact that the relative group size is a random variable does not seem bias the average effect estimation.

Standard error bias is again assessed by subtracting the population standard error value (the standard deviation of all average effect estimators) from the average standard error value and dividing this number by the population standard error value and multiplying by 100. Table 4.4 on the following page gives the results for the three methods.

All three models again yield negligible standard error biases for small,

Table 4.4

Standard Error Biases of the Second MC Study for the Models Including the Interaction Given a ***Moderate*** *Treatment Latent Covariate Correlation.*

Interaction						
0	0.2	0.5	0.8	1.5	3	10
Singlegroup model extended to random slopes						
−1.87	−1.54	−2.32	0.57	−0.94	−1.56	−4.03
Multigroup model with population treatment probability						
−1.85	−2.15	−1.45	1.71	−3.01	−0.13	−3.97
Multigroup model with estimated treatment probability						
−1.99	−2.17	−2.03	0.12	−4.85	−4.15	**−8.39**

Note. Values with an absolute value larger than five are printed bold and are considered severely biased.

medium and large interaction effects. The standard error biases of the single-group model as well as the standard error biases of the multi-group model with the population treatment probability are acceptable for all parameter combinations. The multi-group model using the sample treatment probability yields standard error biases that are acceptable even for very large interaction effects $\beta_3 = 1.5$ and 3. This was not the case in the last Monte Carlo study. The standard error biases of the multi-group model using the sample treatment probability are inflated only for the (very) extreme case $\beta_3 = 10$.

The descriptive statistics of the results of the Monte Carlo study for the

balanced as well as the Monte Carlo study for the unbalanced design show that all three models are applicable to test average treatment effects if small or medium interaction effects are present. Because most interaction effects found in the social sciences are small or medium, the conclusion is that there is usually no need to apply the computationally more demanding SEM approach extended to random slopes. Furthermore, for the given examples, the multiple group approach has proved to be applicable despite the described deviations from normality of several variables. This finding is in line with Bollen (1989, p. 126–127), Johnston (1984, p. 281–285), and Jöreskog (1973, p. 94)).

It is clear however, that the design of the presented Monte Carlo studies is quite small, limiting the generalization of the conclusions drawn from it. An important extension is to include more than two treatment (groups) and a multivariate covariate in order to test, how the outlined procedures perform with regard to analyzing average treatment effects. It was also mentioned that one advantage of the multiple group approach is to model group specific variances of the structural residual. However, the case where the residual variances differ between groups (i. e., $Var(\zeta \mid X{=}1) \neq Var(\zeta \mid X{=}1)$) was not considered in the Monte Carlo studies. It might be possible to extend the random slope approach in order to model different residual variances. Comparing the different approaches with regard to analyzing average treatment effects is interesting because Steyer (2005) developed models to test not only average treatment effects but also individual treatment effects, and these models may imply different residual variances.

Chapter 5

Discussion and Conclusion

In the following section I discuss why I believe that an analysis of average or main effects plays an important role in the analysis of treatment effects, even if interaction effects are present. Given that interaction effects are present, the effect of a treatment depends on covariates. An analysis of these interaction effects therefore provides detailed information. For each (combination of) covariate value(s), it is possible to compute the predicted difference between the treatment groups with respect to the dependent variable. Interaction effects provide information on how the conditional treatment effects vary depending on the covariates.

Some statisticians advocate not to interpret average (or main) effects in the presence of interactions, because an average effect does not adequately represent the variability of the conditional (or simple) effects of the treatment. The analysis of average effects in non-orthogonal designs is even more in dispute, mostly because the partitioning of the sums of squares is still controversial. It is certainly not my intention to argue against the analysis of interaction effects, because they are very informative and because the theoretical focus of most studies is on them rather than on the average effects.

However, even if interaction effects are present, many researchers consider the information about the average effect as informative. This can be seen in the long struggle to find an adequate analysis of non-orthogonal designs (Carlson & Timm, 1974; Gosslee & Lucas, 1965; Keren & Lewis, 1976) and the related debate documented in the Psychological Bulletin. Different perspectives and approaches may be found for example in Snee (1973); M. H. Kutner (1974); Speed et al. (1978). Some of the approaches, especially the four types of partitioning the sums of squares, are widely used and implemented in most statistical software for ANOVA and multiple linear regression. This indicates that there must be substantial interest in the analysis of average effects within the research community and applied fields.

As mentioned before, there are several methodological articles that focus on the analysis of interaction effects *and* average effects of a treatment, especially in the behavioral sciences (see, e. g., Aiken & West, 1991; West et al., 1996). B. O. Muthén and Curran (1997) describe a latent-growth framework to estimate the average effect of a treatment, if the treatment interacts with the initial status, which is the measure of the outcome variable before onset of the treatment.

Average treatment effects also play an important role in the theory of causality (see, e. g., Rubin, 1974) and the analysis of quasi-experimental studies (see, e. g., Shadish et al., 2002) or observational studies (Rosenbaum, 2002). In the literature on average causal effects much effort is spent on developing tests that aim to reduce bias in estimating average treatment effects. In fact, procedures such as the propensity score analysis (Rosenbaum & Rubin, 1984) yield (only) an estimate of the average effect of a treatment (and not of the interaction effects).

The key principle of methods based on the propensity score is to avoid the

modeling of a full regression equation (including interaction terms). Instead, the covariates are used to model (and estimate) the propensity scores (the probability of an observational unit being assigned to one of the treatment conditions). The observational units are then matched based on the (estimated) propensity scores. If certain conditions hold, then the outcome can be interpreted as an estimate of the average causal effect of the treatment (see, e. g., Rosenbaum & Rubin, 1984).

Steyer et al. (2002) discusses conditions that have to hold so that the average treatment effect estimated by the outlined method of this thesis can be interpreted as an unbiased estimate for the average causal effect in the sense of Rubin (1974); Rubin (1978). By applying the outlined method to designs that were traditionally analyzed with ANOVA or multiple linear regression methods, it is possible to compare them to studies analyzed with propensity score methods.

So far it was mentioned that many researchers consider average effects important (at the presence of interaction effects). From a data analysis perspective, average effects may provide a good starting point. Consider a study that investigates the effect of a treatment depending on several categorical and continuous covariates. Even with just a few categorical and continuous covariates, the number of possible effects to analyze becomes quite large. If there is no theory that leads to a-priory hypotheses, it may seem arbitrary which terms of the regression equation to test for significance.

However, a hypothesis about the average treatment effect — no matter how complicated the (underlying) regression equation might be — is always a straightforward hypothesis. Because the degrees of freedom of such a hypothesis are equal to the number of the involved average effects, the power of this test will be larger than the power of tests of more detailed effects. In

this way, a measure of the average treatment effect serves the principle of simplicity.

I certainly agree that average (treatment) effects do not represent the variability of the treatment effects, if interactions with covariates are present. I note however, that conditional effects (or simple effects in ANOVA) which form the interactions also represent averages: the averages of the individual effects of the observational units sharing the same covariate value. It is apparent that interaction effects do not represent the variability of these individual effects. Designs and statistical methods to analyze individual effects are described in the literature on the analysis of individual differences. For example, Steyer (2005) discussed the analysis of individual and average causal effects.

Average effects are also important in applied settings. In all cases where the members of a population can not be treated individually, a decision about which treatment to apply to the population should be based on the average treatment effect. Consider a study of a new educational program. Some students may benefit more from the new program than others. The average effect provides valuable information in order to decide whether to implement the program for a school district or any other population. The example can be transferred to other cases, such as the patients of a clinic or the employees of a firm.

Average effects are important for rankings. Consider a comparison of two clinics with regard to their performance of treating a disorder. The effect of the treatment may depend on a pretest, such as the severity of the disorder or the neediness for a treatment. The average treatment effect provides a useful measure to compare the overall performance of clinic A to clinic B, if the patients of the clinics differ in their covariate values.

5.1 Conclusion

A major improvement in the analysis of treatment effects was the application of multiple linear regression methods that use centered covariates. This allowed to (simultaneously) estimate and test interaction effects as well as average treatment effects for designs including continuous covariates (see, e. g., Aiken & West, 1991; West et al., 1996). In this thesis, I have defined the average treatment effect and shown how multiple linear regression, or to be more specific the general linear model, can be applied to simultaneously analyze interaction and average treatment effects without the need to center the covariates. The method uses the general linear hypothesis and is more flexible than the centering approach especially for the case that more than one covariate is involved and higher order interactions are considered.

An important result of this thesis is that multiple linear regression methods (whether centering is applied or the general linear hypothesis) are only applicable to test average treatment effects, if the means (and higher order moments) of the covariates that appear in the hypotheses about average treatment effects are known. If these covariate means, have to be estimated from the sample, is was argued that maximum likelihood methods should be used to test average treatment effects especially if medium or large interaction effects are present.

Because in many studies it will be the case that the covariate means have to be estimated, the proposed maximum likelihood method is an important improvement for the analysis of treatment effects. I show that the test can be implemented in existing statistical software programs that use maximum likelihood methods, such as LISREL and Mplus. It was also shown that the procedure is applicable to unbalanced designs usually analyzed by non-orthogonal ANOVA methods.

In the social sciences, the covariates are oftentimes measured with measurement error. The appropriate method to address measurement errors of covariates is to apply structural equation modeling (Bollen, 1989). In this thesis, two structural equation modeling approaches are discussed that allow the modeling of interactions between treatment and latent covariates. For the standard multiple group model (section 3.3.1) it was shown how to implement the test for the average treatment effect. Two cases were distinguished depending on whether randomization was implemented in the research design or not. A method to determine power with regard to detecting average treatment effects was described. For the structural equation modeling approach extended to random slopes, which allows the modeling of interactions between continuous (latent) variables, it was also shown how to implement the test of the average treatment effect.

Practical recommendations The Monte Carlo study for the manifest covariate implies that common regression analysis (centering and general linear hypothesis approach) may be used even if the mean of the covariate is only estimated as long as there is only a small interaction between treatment and covariate. If medium or large interaction effects are present, then the maximum likelihood approach should be applied in order to test the average treatment effect.

If covariates are measured with a measurement error, latent variable modeling should be applied (Bollen, 1989). For randomized designs, the standard (multi-group) latent variable framework should be applied with the described restrictions in order to estimate and test average treatment effects. For non-randomized designs the Monte Carlo studies have shown that the SEM approach extended to random slopes as well as the standard multi-

group approach perform equally well for the chosen examples. Only for very large interaction effects did the multi-group approach with the estimated treatment probabilities yield biased standard errors while the random slope approach performed well.

It is clear however, that more testing is required involving larger designs that include cases with multivariate (latent) covariates and more than two treatment groups. For the given examples, however, it can be concluded that the described outline provides a unified approach to estimate and test average (or main) treatment effects, as well as interaction effects for interventional studies. Steyer and Partchev (2007) are developing a user friendly software including the methods described in this dissertation.

Appendix A

Mplus Inputs

```
DATA: FILE IS DATA.dat;
VARIABLE: NAMES ARE Y X Z;
USEVARIABLES ARE Y Z1 Z2 C1 C2 Z1C1 Z2C1 Z1C2 Z2C2;
DEFINE: Z1=0; Z2=0; C1=0; C2=0;
  IF(Z EQ 1)THEN Z1 = 1; IF(Z EQ 2)THEN Z2 = 1;
  IF(X EQ 1)THEN C1 = 1; IF(X EQ 2)THEN C2 = 1;
  Z1C1 = Z1 * C1; Z2C1 = Z2 * C1;
  Z1C2 = Z1 * C2; Z2C2 = Z2 * C2;
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL: Y ON Z1 Z2
        C1(b3)
        C2(b4)
        Z1C1(b5)
        Z2C1(b6)
        Z1C2(b7)
        Z2C2(b8);
  [Z1](mZ1);
  [Z2](mZ2);
  [Y C1 - Z2C2];
MODEL CONSTRAINT:
  b3 = - b5 * mZ1 - b6 * mZ2;
  b4 = - b7 * mZ1 - b8 * mZ2;
```

Listing A.1: Mplus input for the ANOVA example in section 2.5

```
DATA: FILE IS DATA.dat;
VARIABLE: NAMES ARE Y X Z1 Z2;
USEVARIABLES ARE Y Z1 Z2 Z1Z2
  C1 Z1C1 Z2C1 Z1Z2C1
  C2 Z1C2 Z2C2 Z1Z2C2;
DEFINE: Z1Z2 = Z1 * Z2; C1 = 0; C2 = 0;
  IF(X EQ 1)THEN C1 = 1; IF(X EQ 2)THEN C2 = 1;
  Z1C1 = Z1*C1; Z2C1 = Z2*C1; Z1Z2C1 = Z1*Z2*C1;
  Z1C2 = Z1*C2; Z2C2 = Z2*C2; Z1Z2C2 = Z1*Z2*C2;
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL: Y ON Z1 Z2 Z1Z2
           C1(b4)
           Z1C1(b5)
           Z2C1(b6)
           Z1Z2C1(b7)
           C2(b8)
           Z1C2(b9)
           Z2C2(b10)
           Z1Z2C2(b11);
  [Y C1 - Z1Z2C2];
  [Z1](mZ1);
  [Z2](mZ2);
  [Z1Z2](mZ1Z2);
MODEL CONSTRAINT:
  b4 = - b5*mZ1 - b6*mZ2 - b7*mZ1Z2;
  b8 = - b9*mZ1 - b10*mZ2 - b11*mZ1Z2;
```

Listing A.2: Mplus input for the centering example described in section 2.6

```
1 DATA:    FILE = data.dat;
2 VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
3          GROUPING = X (1=G1 2=G2);
4 ANALYSIS: TYPE = MEANSTRUCTURE;
5 MODEL:   XI BY Z1 Z2;
6          ETA BY Y1 Y2;
7          [XI@0];
8          ETA ON XI;
```

Listing A.3: Multi-group SEM to analyze treatment effects for a randomized two-group design modeling an interaction between treatment and a latent covariate (see section 3.4).

```
1  DATA:    FILE = data.dat;
2  VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
3           GROUPING = X (1=G1 2=G2);
4  ANALYSIS: TYPE = MEANSTRUCTURE;
5  MODEL:   XI BY Z1 Z2;
6           ETA BY Y1 Y2;
7           [XI@0];
8           ETA ON XI;
9  MODEL G1: ETA ON XI (B1);
10 MODEL G2: ETA ON XI (B2);
11 MODEL CONSTRAINT:
12          NEW (INT);
13          INT = B2 - B1;
```

Listing A.4: This is the same input as in Listing A.3 with the additional test for the interaction in lines 9 to 13.

```
DATA:    FILE = data.dat;
VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
          GROUPING = X (1=G1 2=G2);
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:   XI BY Z1 Z2;
          ETA BY Y1 Y2;
          [Z1@0];
          [Y1@0];
          ETA ON XI;
MODEL G1: [XI] (MXI);
          [ETA] (AL1);
          ETA ON XI (BE1);
MODEL G2: [XI] (MXI);
          [ETA] (AL2);
          ETA ON XI (BE2);
MODEL CONSTRAINT:
          NEW(AE);
          AE = AL2 - AL1 + (BE2 - BE1) * MXI;
```

Listing A.5: This is an example of an Mplus input to test the average treatment effect. The difference to the Mplus input given in Listing A.3 is the scaling of the latent variables. The test of the average treatment effect requires additional statements (see section 3.4.1).

```
DATA: FILE = pop.dat;
      TYPE = MEANS FULLCOV;
      NGROUPS = 2;
NOBSERVATION = 1000 1000;
VARIABLE: NAMES = Y1 Y2 Z1 Z2;
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
 XI BY Z1-Z2; ETA BY Y1-Y2;
```

```
! XI BY Z1-Z2@1; ETA BY Y1-Y2@1;
 [XI@0];
 ETA ON XI;
! ETA ON XI (1);
! [ETA@0];
OUTPUT: STANDARDIZED RESIDUAL; TECH1;
```

Listing A.6: This input is described throughout section 3.5, discussing power in randomized designs.

```
DATA: FILE = pop.dat;
      TYPE = MEANS FULLCOV;
      NGROUPS = 2;
NOBSERVATION = 1000 1000 ;
VARIABLE: NAMES = Y1 Y2 Z1 Z2;
 USEVARIABLES = Y1 Y2;
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
! ETA by Y1-Y2;
 ETA by Y1-Y2@1;
! [ETA@0];
OUTPUT: STANDARDIZED RESIDUAL; TECH1;
```

Listing A.7: The input for the model ignoring ξ. See section 3.5.4 for a description of this Input.

```
MONTECARLO:
 NAMES ARE Y1 Y2 Z1 Z2; NGROUPS = 2;
 NOBSERVATION = 400 400 ; NREPS = 1000 ;
 SEED = 36555654 ;
 REPSAVE = ALL; SAVE = mcrep*.dat;
MODEL MONTECARLO:
 XI BY Z1-Z2@1; XI@1; Z1-Z2@0.5; [XI@0 Z1-Z2@0];
```

```
 ETA BY Y1-Y2@1; Y1-Y2@0.5; [Y1-Y2@0];
MODEL MONTECARLO-X1:
 ETA ON XI@0.707106781186548;
 ETA@0.5; [ETA@0.2];
MODEL MONTECARLO-X2:
 ETA ON XI@0.907106781186548;
 ETA@0.5; [ETA@0.2];
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
 XI BY Z1-Z2; ETA BY Y1-Y2;
 [XI@0];
 ETA ON XI;
MODEL X1:
 ETA ON XI (p1);
MODEL X2:
 ETA ON XI (p2);
MODEL CONSTRAINT:
NEW (int);
 int = p2 - p1;
```

Listing A.8: The Mplus input for the Monte Carlo study described in section 3.5.2 is given here.

```
MONTECARLO:
 NAMES ARE Y1 Y2 Z1 Z2; NGROUPS = 2;
 NOBSERVATION = 500 500; NREPS = 1000 ;
 SEED = 13155732;
 REPSAVE = ALL; SAVE = mcrep*.dat;
MODEL MONTECARLO:
 XI BY Z1-Z2@1; XI@1; Z1-Z2@0.5; [XI@0 Z1-Z2@0];
 ETA BY Y1-Y2@1; Y1-Y2@0.5; [Y1-Y2@0];
MODEL MONTECARLO-X1:
 ETA ON XI@0; ETA@1; [ETA@0.2];
```

```
MODEL MONTECARLO-X2:
 ETA ON XI@0; ETA@1; [ETA@0.4];
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL:
 XI BY Z1-Z2; ETA BY Y1-Y2;
 [XI@0];
 ETA ON XI;
```

Listing A.9: The Mplus input for the Monte Carlo study described in section 3.5.4.

```
DATA: FILE = mcreplist.dat;
 TYPE = MONTECARLO;
VARIABLE: NAMES = Y1 Y2 Z1 Z2 X;
 GROUPING = X (1 = G1 2 = G2);
 USEVARIABLES = Y1 Y2;
ANALYSIS: TYPE = MEANSTRUCTURE;
MODEL: ETA BY Y1-Y2;
```

Listing A.10: The Mplus input for the Monte Carlo study described in section 3.5.4. This input is for the model without the latent covariate and tau-congeneric measurement models and requires previously generated data (e. g., from the input given in Listing A.9).

Appendix B

R Programs

Listing B.1 gives a program, that computes the power as described in the main text (see, e. g., Chapter 3.5). This program is written for the statistical language and programming environment R (R Development Core Team, 2006). The values of *chi.square.diff* have to be obtained as described in the text for a total sample size *samplesize.orignial*. Given *alpha*, *df*, and *samplesize.desired* the program computes the power.

```
1 q <- qchisq(1 - alpha, df)
2 power <- 1 - pchisq(q, df, ncp = chi.square.diff *
3  samplesize.desired / samplesize.original)
```

Listing B.1: A R program to compute power.

```
mc.sample <- function(filename) {
  xi <- rnorm(N, mean=0, sd=1)
  x <- 1 + (runif(N) <= 1 / (1 + exp(l0 + l1*xi)))
  eta <- 0 + 0 * (x==2) + sqrt(0.5) * xi +
    b3 * xi * (x==2) + rnorm(N, 0, sqrt(0.5))
  z <- rep(1, N) %o% c(0,0) + xi %o% c(1,1) +
    cbind(rnorm(N, 0, sqrt(0.5)), rnorm(N, 0, sqrt(0.5)))
  y <- rep(1, N) %o% c(0,0) + drop(eta) %o% c(1,1) +
    cbind(rnorm(N, 0, sqrt(0.5)), rnorm(N, 0, sqrt(0.5)))
  write.table(cbind(y, z, x), file = filename, sep = ",",
              row.names=FALSE, col.names=FALSE,
              quote=FALSE)
}
writeLines(paste("mcrep", 1:1000, ".dat", sep = ""),
           "mcreplist.dat", "\n")
set.seed(310001317)
lapply(readLines("mcreplist.dat"), function(fn)
        mc.sample(fn))
```

Listing B.2: R input to generate data for Monte Carlo studies. The assignment probabilities are computed by a logistic function of the latent covariate.

Appendix C

Proofs

C.1 Ignoring the Covariate

Proof of Equations 2.17 and 2.18:

The regression of Equation 2.1 is repeated here:

$$E(Y \mid X, Z) = \beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX \tag{C.1}$$

To compute the two group means of Y we need the regression of Y on X. The following derivation uses some properties of regression and conditional expected values, which can be found for example in Bauer (1981).

$$\begin{aligned} E(Y \mid X) &= E\left[E(Y \mid X, Z) \mid X\right] \\ &= E\left[\beta_0 + \beta_1 Z + \beta_2 X + \beta_3 ZX \mid X\right] \\ &= \beta_0 + \beta_1 E(Z \mid X) + \beta_2 X + \beta_3 E(ZX \mid X) \\ &= \beta_0 + \beta_1 E(Z \mid X) + \beta_2 X + \beta_3 X E(Z \mid X) \end{aligned} \tag{C.2}$$

The two group means of Y result in:

$$\begin{aligned} E(Y \mid X{=}0) &= \beta_0 + \beta_1 E(Z \mid X{=}0) \\ E(Y \mid X{=}1) &= \beta_0 + \beta_1 E(Z \mid X{=}1) + \beta_2 + \beta_3 E(Z \mid X{=}1). \end{aligned} \tag{C.3}$$

C.2 Power for Randomized Designs

The following proof shows that, given the parameter settings in section 3.5.1, the variance of the latent outcome is always one (see Equation 3.37 on page 74):

Proof.

$$\begin{aligned} Var^{(1)}(\eta) &= Var^{(1)}(E(\eta \mid \xi)) + Var^{(1)}(\zeta) \\ &= Var^{(1)}(\alpha^{(1)} + \beta^{(1)}\xi) + Var^{(1)}(\zeta) \\ &= {\beta^{(1)}}^2 \, Var^{(1)}(\xi) + Var^{(1)}(\zeta) \\ &= \frac{1 - Var^{(1)}(\zeta)}{Var^{(1)}(\xi)} \, Var^{(1)}(\xi) + Var^{(1)}(\zeta) \\ &= 1. \end{aligned} \tag{C.4}$$

□

As a consequence the correlation of ξ and η is equal to the structural slope in the first group.

Proof.

$$\begin{aligned} Corr^{(1)}(\xi, \eta) &= \frac{Cov^{(1)}(\xi, \eta)}{Std(\xi) \, Std^{(1)}(\eta)} \\ &= Cov^{(1)}(\xi, \alpha^{(1)} + \beta^{(1)}\xi + \zeta) \\ &= \beta^{(1)} \, Var(\xi) \\ &= \beta^{(1)} \end{aligned} \tag{C.5}$$

□

Bibliography

Aiken, L. S., West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage. 5, 10, 34, 54, 77, 135, 138

Angrist, J. D., Imbens, G. W., Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. 13

Asparouhov, T., Muthén, B. O. (2002). *Full–information maximum–likelihood estimation of general two–level latent variable models.* (In preparation) 108, 112

Baron, R. M., Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173-1182. 3, 12, 55

Bauer, H. (1981). *Probability theory and elements of measure theory.* New York: Academic Press. 150

Bentler, P. M., Chou, C. P. (1988). Practical issues in structural equation modeling. In J. S. Long (Ed.), *Common problems/proper solutions: Avoiding error in quantitative research.* Newbury Park, CA: Sage. 72

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley. 50, 57, 119, 133, 139

Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, *61*, 109-121. 107

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, *53*, 605-634. 73

Carlson, J. E., Timm, N. H. (1974). Analysis of nonorthogonal fixed–effects designs. *Psychological Bulletin*, *81*, 563-570. 30, 135

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, *13*, 261–281. 5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Erlbaum. 27, 40, 45, 71, 75, 122

Cohen, J., Cohen, P., West, S. G., Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahawah, NJ: Erlbaum. 4, 20, 31, 50

Cook, T. D., Campbell, D. C. (1979). *Quasi–Experimentation*. Chicago: Rand McNally. 4

Cox, D. R., Hinkley, C. V. (1974). *Theoretical Statistics*. London: Chapman & Hall. 27, 30, 47

Cronbach, L. J., Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington. 4

Difference Testing Using the Loglikelihood. (n.d.). Retrieved November 17, 2006, from `http://www.statmodel.com/chidiff.shtml`. 113

Draper, N., Smith, H. (1981). *Applied regression analysis*. New York: Wiley. 20

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage. 15

Gelman, A., Meng, X.-L. (2004). *Applied bayesian modeling and causal inference from incomplete–data perspectives*. Chichester: Wiley. 21

Gosslee, D. G., Lucas, H. L. (1965). Analysis of variance of disproportionate data when interaction is present. *Biometrics*, *21*, 115-133. 4, 30, 135

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., Ostrowski, E. (Eds.). (1994). *Handbook of Small Data Sets*. London, UK: Chapman & Hall. 42

Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association*, *81*, 945–970. 17

Jaccard, J., Turrisi, R., Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage. 10, 54

Jaccard, J., Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage. 8, 62, 65, 66, 68, 107

Johnston, J. (1984). *Econometric Methods*. New York: McRaw-Hill. 119, 133

Jöreskog, K. G. (1973). A General Method for Estimating a Linear Structural Equation System. In A. S. Goldberger O. D. Duncan (Eds.), *Structural Equation Models in the Social Sciences*. New York: Seminar Press. 119, 133

Jöreskog, K. G., Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books. 57, 59, 60

Jöreskog, K. G., Sörbom, D. (1996). *LISREL 8 user's reference guide*. Chicago, Il: SSI (Scientific Software International). 26

Jöreskog, K. G., Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. Marcoulides R. Schumaker (Eds.), *Advanced Structural Equation Modeling* (p. 57-81). Mahwah, N. J.: Lawrence Erlbaum Associates. 107

Judd, C. M., McClelland, G. H. (1989). *Data analysis: A model-comparison approach*. San Diego: Harcourt Brace Jovanovich. 10, 54

Keppel, G., Zedeck, S. (2000). *Data analysis for research designs: analysis of variance and multiple regression/correlation approaches* (7th ed.). New York: Freeman. 4

Keren, G., Lewis, C. (1976). Nonorthogonal designs: Sample versus population. *Psychological Bulletin*, *83*, 817-826. 30, 135

Klein, A., Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457-474. 108, 112

Kutner, M., Nachtsheim, C., Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). Columbus, OH: McGraw-Hill. 20

Kutner, M. H. (1974). Hypothesis testing in linear models. *The American Statistician*, *29*, 98-100, 133-134. 5, 135

Little, R. C., Freund, R. J., Spector, P. (1991). *SAS System for Linear Models* (3rd ed.). Cary, NC: SAS Institute Inc. 4, 30

Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, *75*, 87-91. 5, 34

Moosbrugger, H. (1981). Zur differentiellen Validität bei nichtlinearen Test-Kriterium-Zusammenhängen. *Zeitschrift fr Differentielle und Diagnostische Psychologie*, *2*, 219-274. 4

Muthén, B. O., Asparouhow, T. (n.d.). *Modeling interactions between latent and observed continuous variables using maximum-likelihood estimation in Mplus. Mplus Web Note #6. Version 1.* Retrieved November 10, 2006, from `http://www.statmodel.com/examples/webnote.shtml`. 108

Muthén, B. O., Curran, P. J. (1997). General Longitudinal Modeling of Individual Differences in Experimental Designs: A Latent Variable Frame-

work for Analisys and Power Estimation. *Psychological Methods*, *2*(4), 371-402. 5, 77, 135

Muthén, L., Muthén, B. O. (2004). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén. 26, 30, 40, 112

Nagengast, B. (2006). *Standard Errors of ACE Estimates: Comparing adjusted group means against the adjusted grand mean. A simulation study*. Unpublished master's thesis, Friedrich Schiller University, Germany. 107

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*, *5*, 465–472. (Section 9. Reprint 1990) 7

Overall, J. E., Spiegel, D. k. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, *79*, 311-322. 2, 5, 30

Overall, J. E., Spiegel, D. K., Cohen, J. (1975). Equivalence of orthogonal and nonorthogonal analysis of variance. *Psychological Bulletin*, *82*, 182-186. 2, 5, 30

R Development Core Team. (2006). R: A language and environment for statistical computing [Computer software and manual]. Retrieved from `http://www.R-project.org`. Vienna, Austria: R Foundation for Statistical Computing. 42, 115, 148

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer. 1, 6, 13, 33, 135

Rosenbaum, P. R., Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524. 6, 135, 136

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5),

688–701. 6, 7, 21, 135, 136

Rubin, D. B. (1978). Bayesian-inference for causal effects - role of randomization. *Annals of Statistics*, *6*(1), 34–58. 21, 136

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. New York: Cambridge University Press. 13, 21

Saris, W. E., Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen J. S. Long (Eds.), *Testing structural equation models.* Newbury Park, CA: Sage. 28, 69, 70

Saris, W. E., Stronkhorst, L. H. (1984). *Causal modeling in nonexperimental research: An introduction to the LISREL approach*. Amsterdam: Sociometric Research Foundation. 28, 69, 70

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In R. Heijmans, D. Pollock, A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (p. 233-247). London: Kluwer Academic Publishers. 113

Satorra, A., Saris, W. (1985). Power of the likelihood ratio test in covariance struture analysis. *Psychometrika*, *51*, 83-90. 28, 69, 70

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement*, *16*, 209-222. 3, 10, 12, 54

Searle, S. R. (1971). *Linear Models*. New York: Wiley. 15

Searle, S. R. (1987). *Linear models for unbalanced data*. New York: John Wiley. 4, 30

Searle, S. R., Casella, G., McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley. 5, 30

Shadish, W. R., Cook, T. D., Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston,

MA: Houghton Mifflin Co. 6, 52, 135

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, *13*, 238-241. 17

Snee, R. D. (1973). Some aspects of nonorthogonal data analysis. *Journal of Quality Technology*, *5*(2), 67-79. 5, 135

Sörbom, D. (1978). An Alternative to the Methodology for Analysis of Covariance. *Psychometrika*, *43*(3), 381-396. 57, 59

Sörbom, D. (1982). Structural equation models with structured means. In K. G. Jöreskog H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction* (p. 183-195). Amsterdam: North Holland. 59

Speed, F. M., Hocking, R., Hacknew, O. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, *73*, 105-112. 5, 135

Steyer, R. (2001). Classical Test Theory. In C. Ragin T. Cook (Eds.), *International Encyclopedia of the Social and Behavioural Sciences. Logic of Inquiry and Research Design* (p. 481-520). Oxford: Pergamon. 53

Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 39-54. 110, 133, 137

Steyer, R., Eid, M. (2001). *Messen und Testen* [Measuring and testing] (2nd ed.). Heidelberg, Germany: Springer. 53

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C. (2000). Causal Regression Model II: Unconfoundedness and Causal Unbiasedness. *Methods of Psychological Research Online*, *5*(3), 55-86. 13

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional aver-

age causal effects. *Methods of Psychological Research Online*, *7*, 41–68. 13, 21, 136

Steyer, R., Partchev, I. (2007). EffectLite: User's Manual. A Program for the Uni- and Multivariate Analysis of Unconditional, Conditional and Average Mean Differences Between Groups [Computer software and manual]. Retrieved from `http://www.metheval.uni-jena.de/projekte/statlite/`. Jena, Germany: University of Jena. 140

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., Fliege, C. (2007). *Causal Effects in Between-Group Experiments and Quasi-Experiments: Theory.* Manuscript in preparation, Friedrich Schiller University Jena, Germany. 12, 13, 33, 57, 68, 115

Stuard, A., Ord, K. J. (1994). *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory* (6th ed., Vol. 1). New York: Oxford Universtiy Press. 27, 30, 47

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference.* New York: Springer. 30, 47

West, S. G., Aiken, L. S., Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, *64*(1), 1-48. 5, 11, 34, 36, 135, 138

Wüthrich-Martone, O. (2001). *Causal modeling in psychology with qualitative independent variables.* Aachen: Shaker. 30

Ehrenwörtliche Erklärung

Mir ist die geltende Promotionsordnung bekannt. Ich habe die Dissertation selbst angefertigt, insbesondere die Hilfe eines Promotionsberaters nicht in Anspruch genommen, und alle von mir benutzten Hilfsmittel und Quellen in meiner Arbeit angegeben. Mir hat niemand bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes unterstützt (entgeltlich/unentgeldlich). Niemand hat darüber hinaus weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Ich habe die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Ich habe weder die gleiche, eine in wesentlichen Teilen ähnliche noch eine andere Abhandlung bei einer anderen Hochschule bzw. anderen Fakultät als Dissertation eingereicht. Ich versichere, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Pittsburgh, Pennsylvania, 10. September 2007

Unterschrift